POLS/CSSS 503:
Advanced Quantitative Political Methodology

Inference and Interpretation of Linear Regression

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Statistical inference

Estimating the linear relationship between x and y is useful

But how certain are we of these estimates?

Even assuming

- linearity of the relationship between x and y,
- normality of $\mathrm{E}(y|x)$
- independence and equal variance of each observation's error term
- no correlation between ε and x

are we sure we got the "right" estimate?

If we sampled more data, would we get the same estimates?

How close to the truth would our estimates be on average?

Statistical inference

If we sampled more data, would our estimator produce the same estimates?

Usually, we can't answer this question with certainty

But we can **quantify our uncertainty** regarding $\hat{\beta}_0$, $\hat{\beta}_1$, etc.

Our goal as quantitative scientists:

Estimating unknowns and **quantifying the uncertainty of those estimates**

Estimates without measures of uncertainty aren't trustworthy or useful

Statistical inference

Where we are headed:

uncertainty of regression results

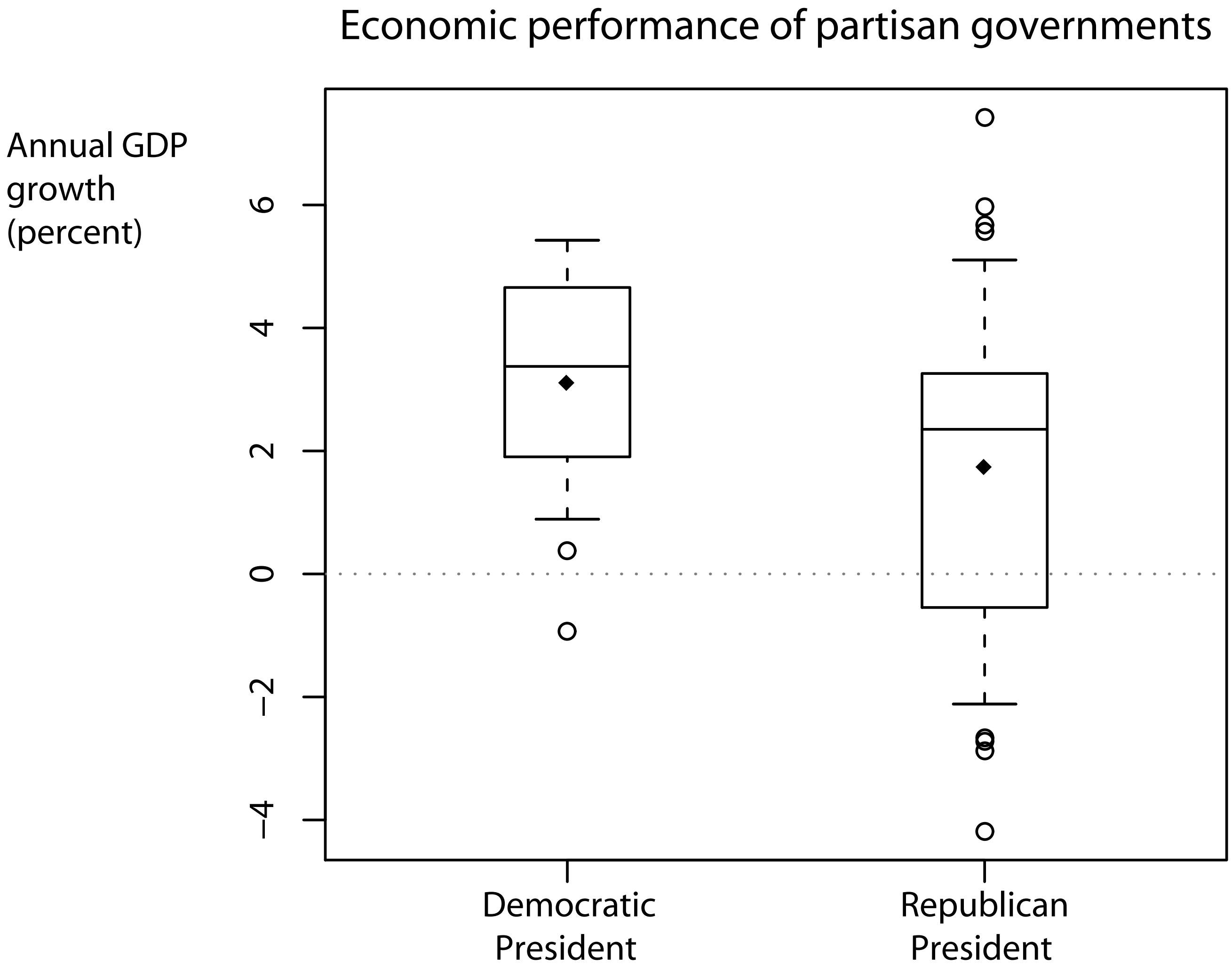
Start simpler:

uncertainty of difference in two means

Example:

Do Democratic presidents enjoy statistically significantly higher growth?

Difference between these means statistically meaningful?



Note difference between this question and the issue of substantive significance

Comparison of two means

How do we quantify certainty of estimates exactly?

Recall that to estimate the mean of a population from a sample we calculate

$$\mathrm{E}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It follows that the difference of two means can be estimated by

$$\mathrm{E}(x-y) \quad = \mathrm{E}(x) - \mathrm{E}(y)$$

Comparison of two means

How do we quantify certainty of estimates exactly?

Recall that to estimate the mean of a population from a sample we calculate

$$\mathrm{E}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It follows that the difference of two means can be estimated by

$$\mathrm{E}(x-y) \quad = \mathrm{E}(x) - \mathrm{E}(y) \quad = \frac{1}{n}\sum_{i=1}^{n} x_i - y_i$$

If we n contained all the population, we'd be done

(What is the population here?)

If n is a sample, we can't be certain the sample is representative "enough" to produce the an accurate estimate of $\mathrm{E}(x-y)$

Comparison of two means

Suppose the sampling was done at random.

Then for each sample, we get a different estimate of $\mathrm{E}(x - y)$

If we could describe the distribution of estimates of $\mathrm{E}(x - y)$,

we could say something about how likely it is (in principle) that

estimates of $\mathrm{E}(x - y)$ take on values in a certain range

this helps us quantify how certain we are that *one specific estimate* of $\mathrm{E}(x - y)$ is close to the true difference

To get there, we need to introduce some new distributions and concepts

The χ^2 distribution

What if we have a variable that is the sum of n squared independent standard Normal RVs

$$X^2 = x_1^2 + x_2^2 + \ldots x_n^2$$

→ Sum of a finite set of Normal RVs, so the Normal only applies approximately

What distribution does this sum really follow?

The χ^2 (chi-squared) distribution,

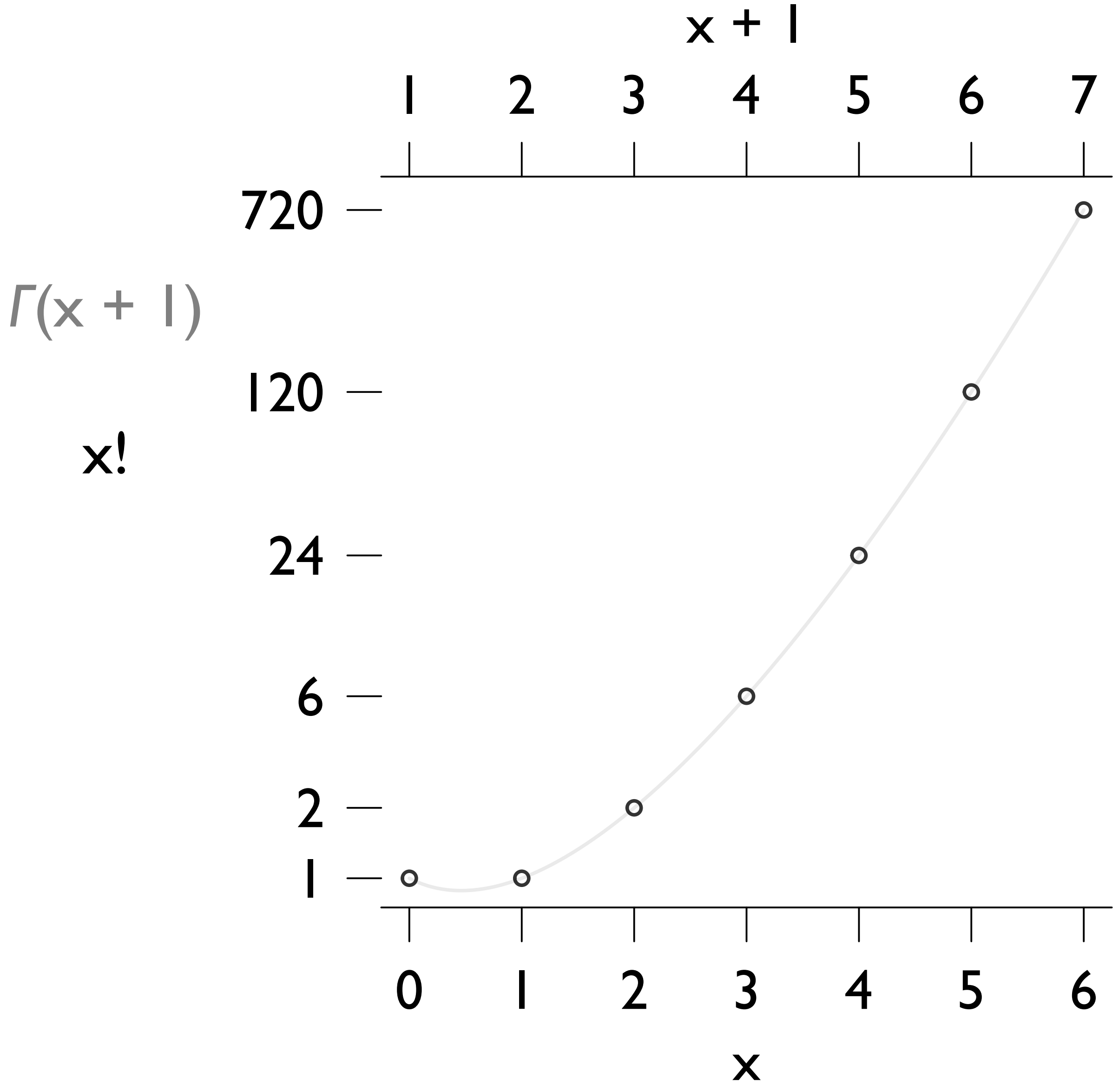
$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(x^2)^{(n-2)/2}\exp(-x/2)$$

which has “degrees of freedom” n

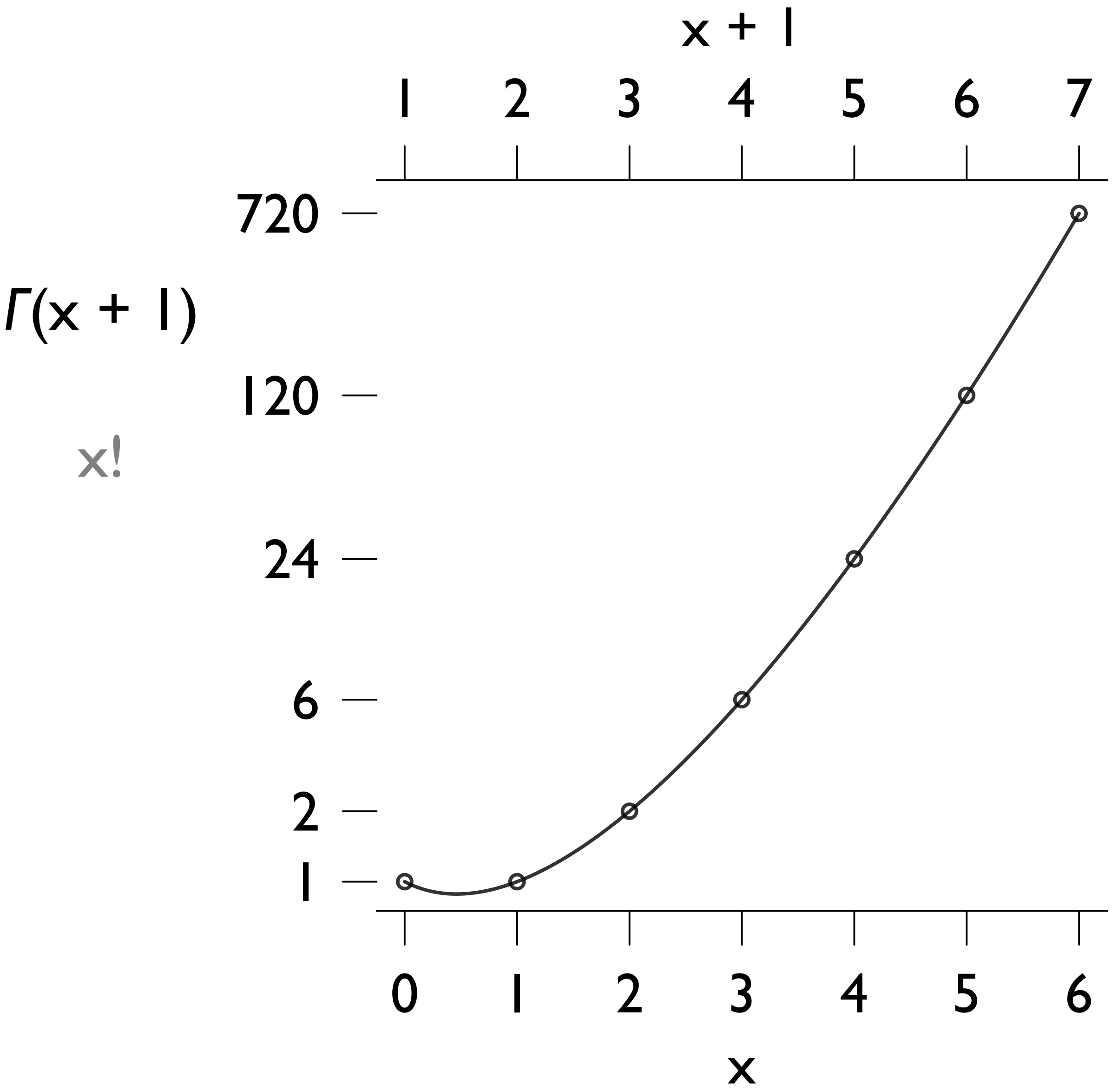
$\Gamma(\cdot)$ is the Gamma function, an interpolated factorial (Aside. . .)

$\mathrm{E}(\chi^2) = n$ and $\mathrm{Var}(\chi^2) = 2n$

Aside: Factorials of x

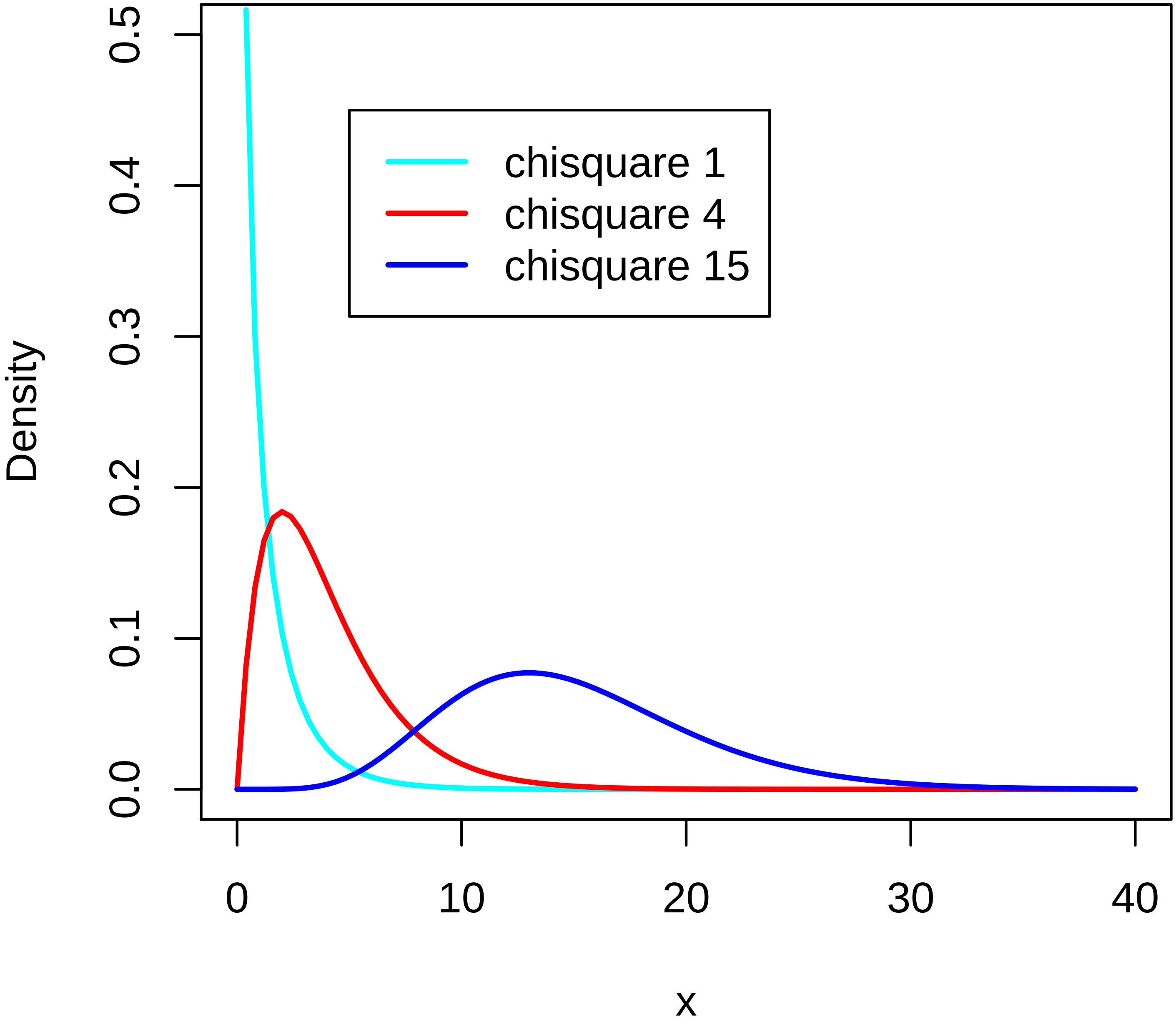


Aside: Γ(x+1)



NB: It's computationally easier to work with $\log(\Gamma(x))$. In R, use `lgamma()`

χ^2 distributions (from Kevin Quinn)



The t distribution

The χ_2 is a key building block for an even more useful distribution

Suppose Z is standard normal and X^2 is distributed χ^2 with n degrees of freedom.

Define

$$t = \frac{Z}{\sqrt{X^2/n}}$$

which is distributed t with n degrees of freedom:

$$f_t(t,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \times (1 + \frac{t^2}{n})^{-\frac{n+1}{2}}$$

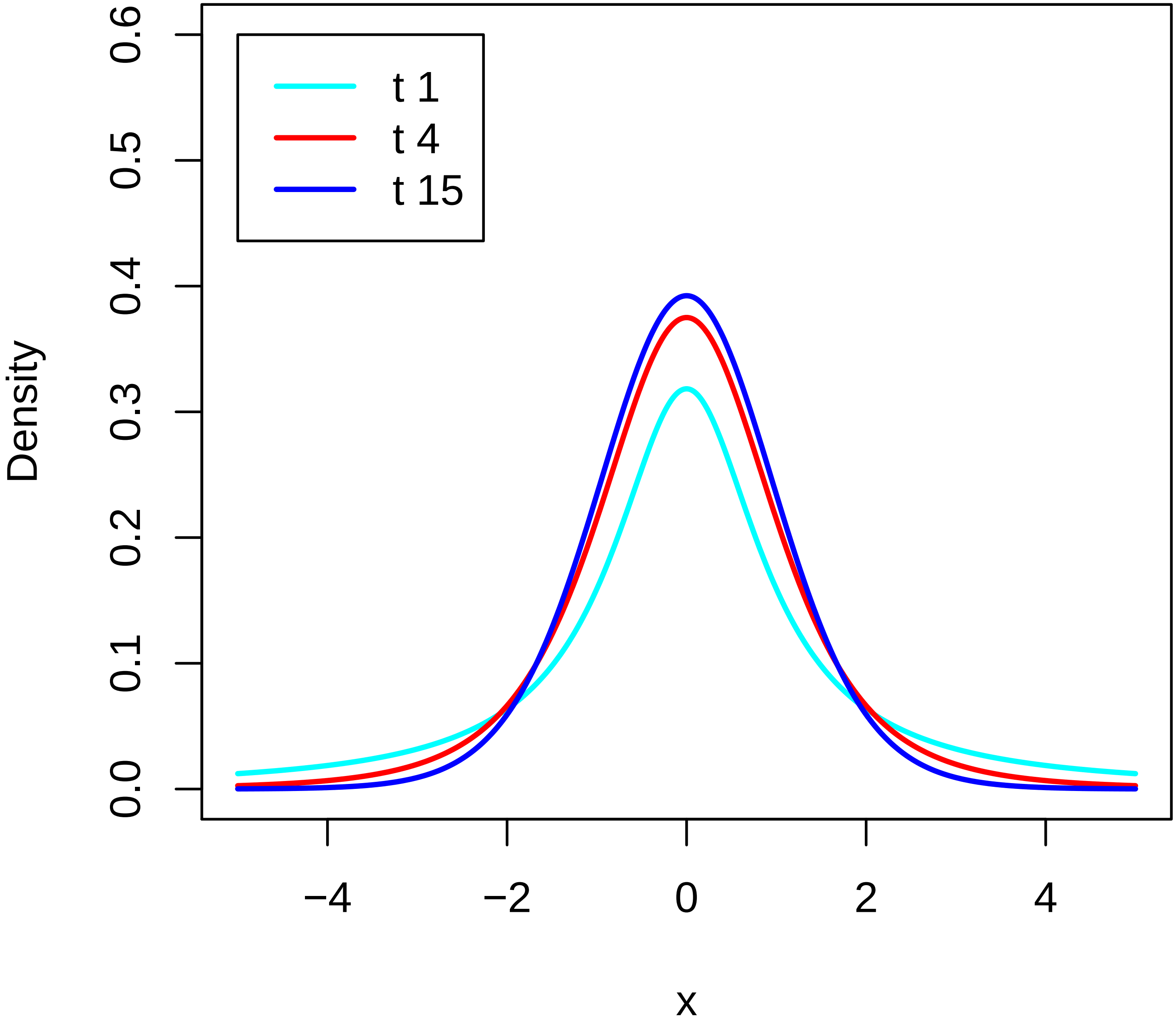
$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ for $n > 2$. Not defined for $n = 1$.

As the degrees of freedom grow, the t distribution approximates the Normal

For low degrees of freedom, the t has fatter tails

t distributions (from Kevin Quinn)



Uses of the t distribution

The t distribution has many uses and an amusing origin (see William Gosset)

Two key uses:

Substitute for normal distribution when extreme errors may be common

Significance tests (today's use)

Start with a simple example:
Suppose I start passing you numbers which I claim come from a certain distribution, but each number is very far from the mean

At what point do you suspect that I am lying about the origin of these numbers, such that they probably come from some other distribution?

Could you quantify the probability I am lying
given the extremity of the numbers?

Uses of the t distribution

More formally. . .

Suppose we have a variable t that is t-distributed with mean 0 and 5 dfs

That is, $\mathrm{P}(t) = f_t(5)$

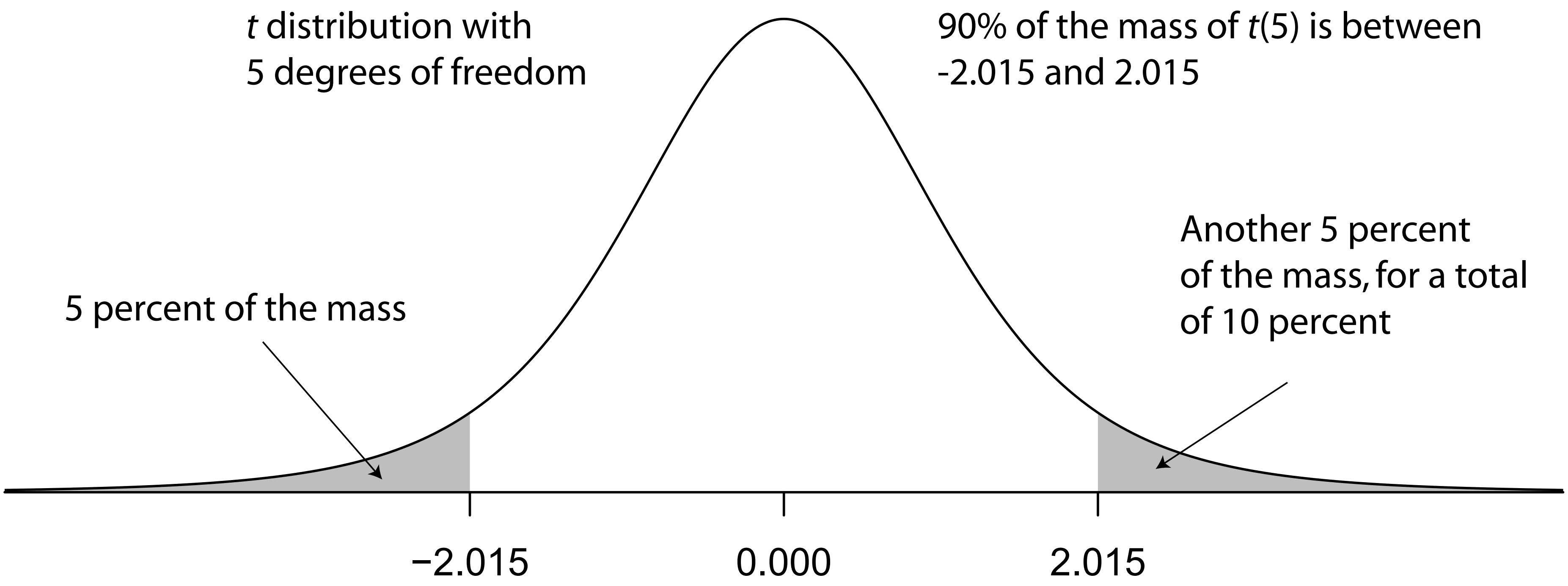
How large would t need to be for us to doubt it came from this distribution?

Put another way, what are the "critical" values of t we would see just

- once in 10 draws?
- once in 20 draws?
- once in 100 draws?

Put still another way,
which critical values will bound the 90% (or 95%, or 99%) most ordinary t draws?

Areas under the t



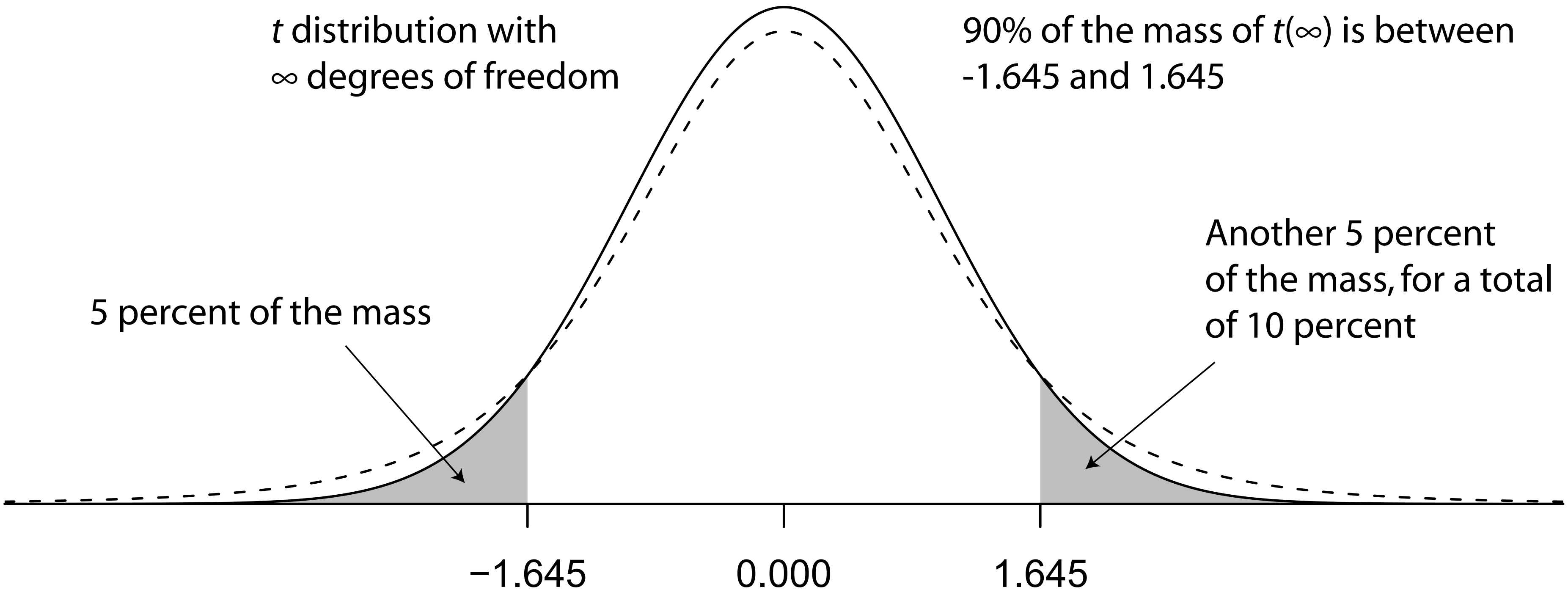
A unusual value is one in the tails. Critical values $=$ cutoff for “unusualness”

To get the curve: `dt(x,df)`

To get critical values: `qt(quantile,df)`. Here, quantiles are 0.05 & 0.95

To get areas of shaded regions: `pt(critical,df)` and `1-pt(critical,df)`

Areas under the t



t distribution with ∞ degrees of freedom

90% of the mass of *t*(∞) is between -1.645 and 1.645

5 percent of the mass

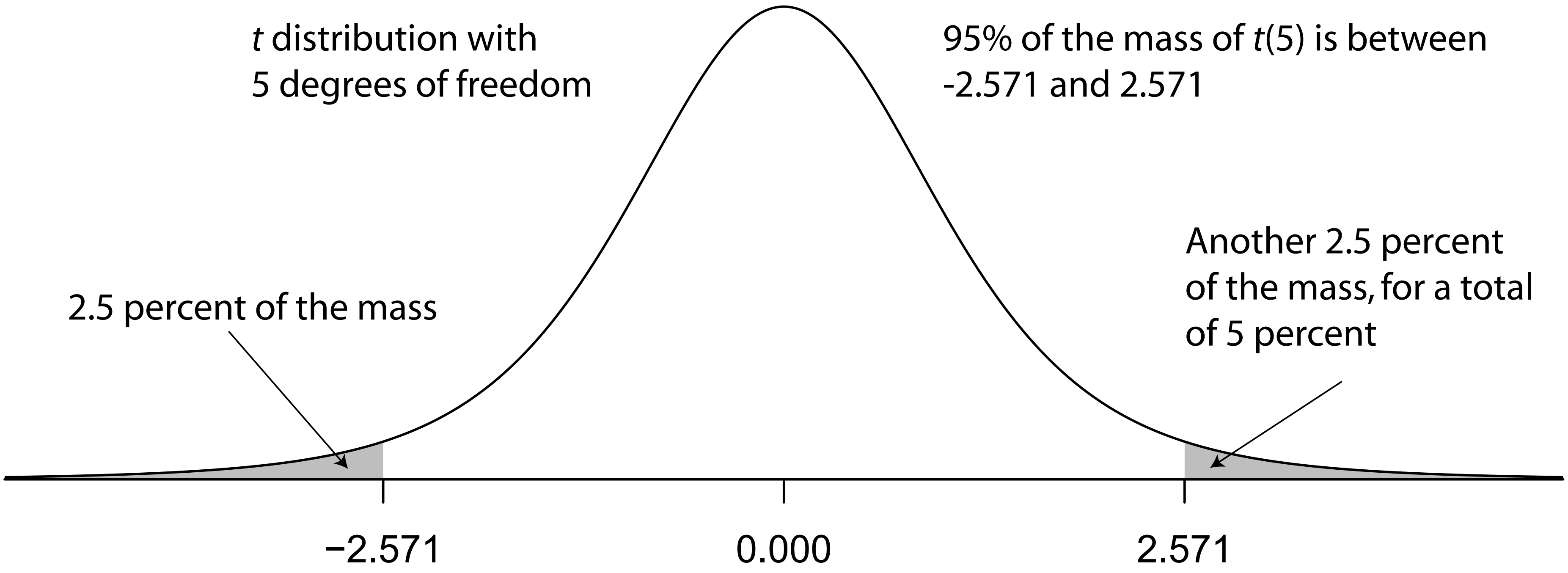
Another 5 percent of the mass, for a total of 10 percent

The degrees of freedom reflect how much information we have

More information makes the tails thinner

More info $\rightarrow$ Critical values shrink $\rightarrow$ estimates get more certain

Areas under the t

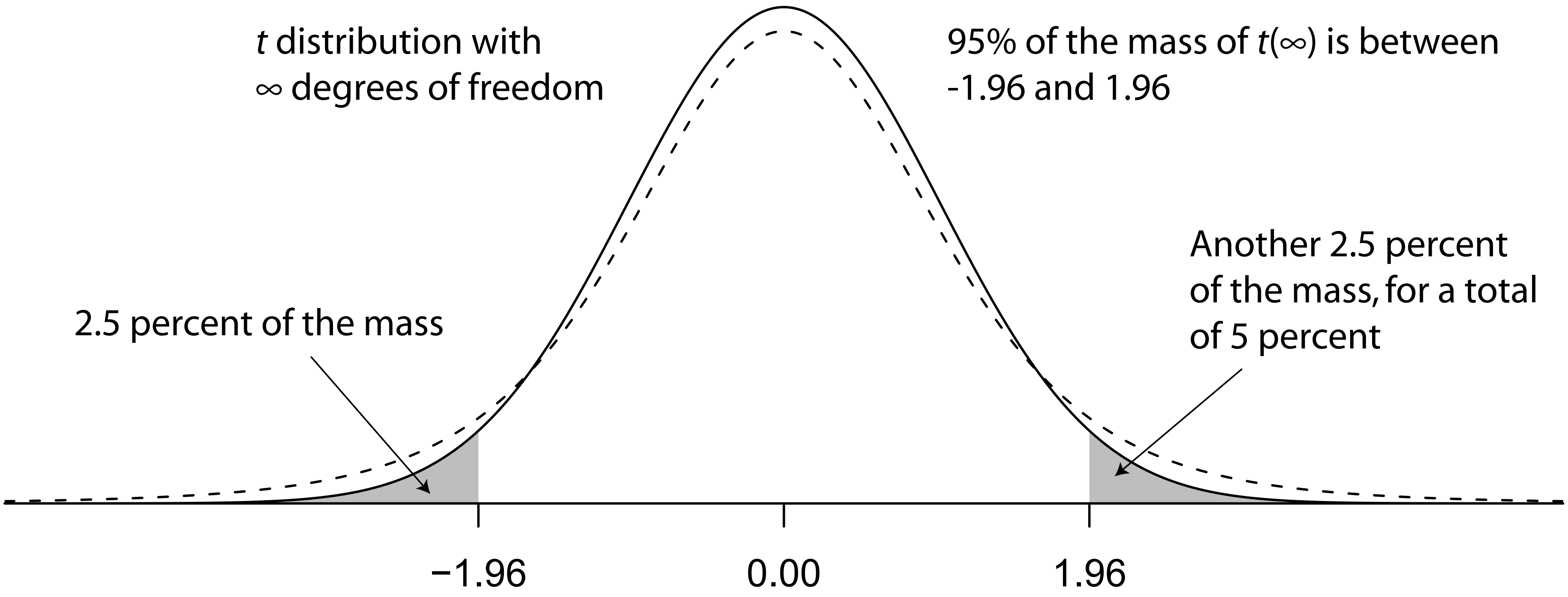


Going back to the $df = 5$ case, notice we can choose what constitutes unusual

Here, we've raise the bar: only the 5% most extreme values are unusual

Because values could be extremely small or extremely large,
each tail thus includes 2.5% of the total probability

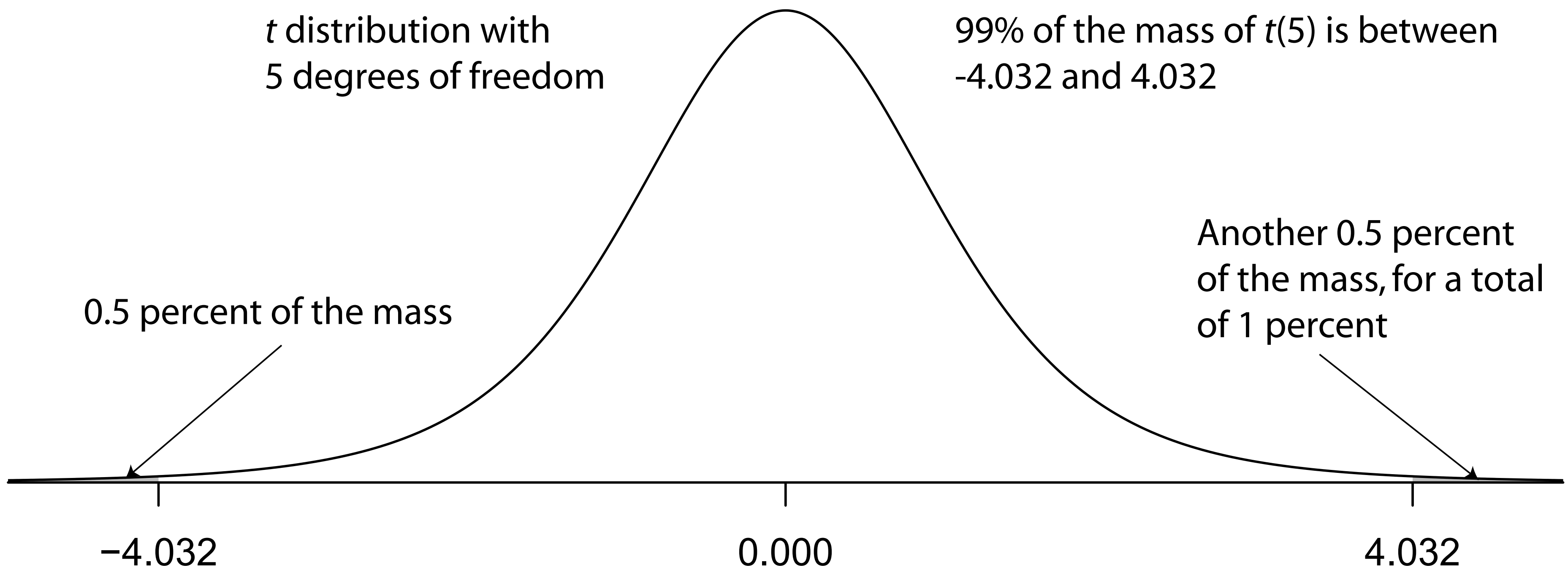
Areas under the t



The infinite degrees of freedom critical values for the 95% case

This is the most widely used standard for whether a result is unusual

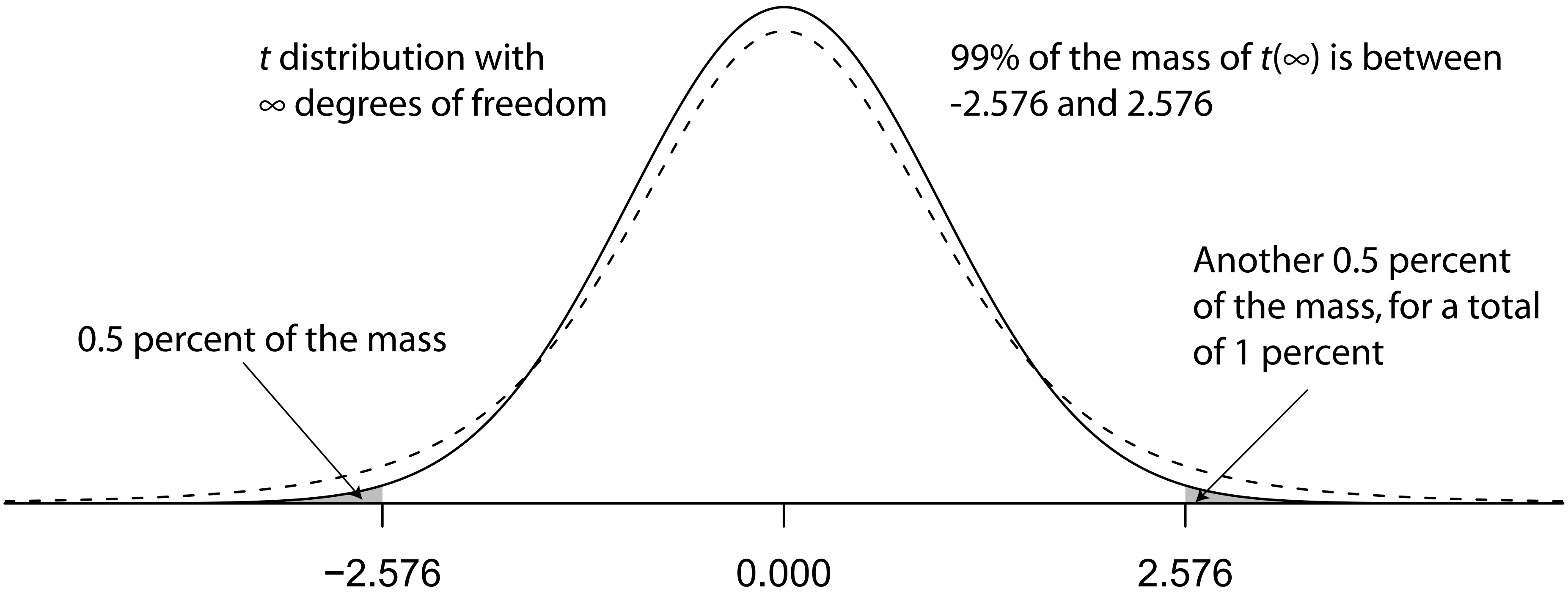
Areas under the t



The most stringent standard is 99%

In this case, a draw from the t must be in the 1% most extreme region to be considered unusual

Areas under the t



The infinite degrees of freedom case for 99%

Critical values of the t distribution

We can state how unusual a value of t is under the assumption that the value of t actually follows a $t(n)$ distribution

Test level	$\text{df} = 5$	$\text{df} = \infty$
0.1 level / 90%	2.015	1.645
0.05 level / 95%	2.571	1.960
0.01 level / 99%	4.032	2.576

These will be very useful for quantifying the uncertainty of estimates

I realize I haven't told you how the t distribution is used yet

For now, just remember we have a tool for measuring how usually a number is if if *really* follows a t distribution

Ways to summarize uncertainty

Now that we have some statistical foundations, what will we build on them?

There are two main ways to summarize uncertainty:

- **Confidence intervals**
- **Significance tests**

I (and many other methodologists) prefer confidence intervals

- Somewhat easier to interpret
- Somewhat less likely to mislead (though not foolproof)

Significance tests standard approach for many decades & remain common

You need to understand both ways to show uncertainty

Significance tests

Suppose we wanted to compare the "unusualness" of two possible values of $\bar{x}$.

We might wonder:
if the population $\bar{x}$ is 0, how unlikely would it be to observe a sample $\bar{x} = 1.89$?

We can calculate this probability, under the assumption that $\bar{x}/\mathrm{se}(\bar{x})$ is distributed t.

$\mathrm{se}(\bar{x})$ is the standard error of the (estimated) mean

That is,
how much we would expect to miss the truth by on average with repeated sampling?

Why assume $\bar{x}/\mathrm{se}(\bar{x})$ is t distributed, and not, say, Normally distributed?

Because the form the standard error takes is $\mathrm{se}(\bar{x}) = \sqrt{\sigma^2/n}$

This matches the denominator of $t = Z/\sqrt{X^2/N}$,
and only approximates the Normal for large N

Significance tests

Neyman-Pearson hypothesis testing in general:

- Set up a baseline, or **null** hypothesis for value of $\bar{x}$
- Calculate the prob of seeing a $\bar{x}$ as distant from the **null** as you did (this is the "p-value")
- If that probability is above a pre-committed threshold, "reject" the alternative hypothesis (your finding) in favor of the null

Notice the implicit notion that you could sample more data

Frequentist inference: emphasis on data you could (but didn't) sample

One of two main branches of inference (the other is Bayesian)

Significance tests

Neyman-Pearson hypothesis testing as usually applied

- Set **null** hypothesis to be $\bar{x} = 0$
- Calculate the t-statistic, $\bar{x}/\mathrm{se}(\bar{x})$
- Look up the p-value of seeing a t-statistic as large as you did given the dfs
  In `R`, `2*pt(abs(tstat),df,lower.tail=FALSE)`

  (Why this formula? Why `2*`?)
- If that p-value is above 0.05, "reject" the finding in favor of the null
- Conversely, if p is below 0.05,
  call the estimate "significantly different from zero at the 0.05 level"

In English: if all our assumptions are correct, sampling error will only falsely produce an effect different from zero 1 out of 20 times that we conduct a t-test

Significance tests

Type I error: Probability of falsely rejecting the null

Type II error: Probability of falsely accepting the null

Significance tests minimize Type I error at the expense of Type II

Seen by some as “conservative”

My view: conservative is another way of saying “wrong in a certain direction,” which is another way of saying biased.

Who privileged the null over the alternative?

We chose both null and alternative hypotheses—
why should we want to treat them asymetrically?

Does it make sense to privilege your “second-best” hypothesis?

→ Reporting confidence intervals rather than p-values avoids null hypotheses and helps focus on the substance of your result

Significance test for a single mean

Generally, t-statistics are the ratio of the estimate to its standard error

In this case,

$$t = \frac{\bar{x}}{\mathrm{se}(\bar{x})}$$

The standard error of the mean is

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n_x}} = \frac{\sigma_x}{\sqrt{n_x}}$$

Notice this gets smaller the more data we have

t can be compared to the critical value of t with $n - 1$ degrees of freedom for a significance level, α (e.g., $\alpha = 0.05$)

Or we can calculate the probability of getting so extreme a t from a random draw from the t distribution with $n - 1$ degrees of freedom

Wait! What are degrees of freedom (df)?

Degrees of freedom:
The number of separate pieces of information used to calculate a statistic

“separate” $=$ “freely movable”

Not the same thing as the number of observations (may be the same as N or less)

Relevance: how many quantities could we estimate from a set of data?

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

two numbers, x_1 and x_2 2 pieces

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

two numbers, x_1 and 3 1 piece

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

two numbers, x_1 and x_2 2 pieces

Now I decide to add an assumption regarding the value of x_2

two numbers, x_1 and 3 1 piece

Instead, suppose I assume I know the mean?

two numbers, x_1 and x_2, and $\bar{x} = 2$ 1 piece

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

two numbers, x_1 and x_2	2 pieces

Now I decide to add an assumption regarding the value of x_2

two numbers, x_1 and 3	1 piece

Instead, suppose I assume I know the mean?

two numbers, x_1 and x_2, and $\bar{x} = 2$	1 piece

How does this work at larger scales?

fifty numbers, $x_1, \ldots x_{50}$, and $\bar{x} = 2$	49 pieces

What are degrees of freedom (df)?

How many separate pieces of unspecified information to estimate?

two numbers, x_1 and x_2	2 pieces

Now I decide to add an assumption regarding the value of x_2

two numbers, x_1 and 3	1 piece

Instead, suppose I assume I know the mean?

two numbers, x_1 and x_2, and $\bar{x} = 2$	1 piece

How does this work at larger scales?

fifty numbers, $x_1, \ldots x_{50}$, and $\bar{x} = 2$	49 pieces
fifty numbers, $x_1, \ldots x_{50}$, $\bar{x} = 2$, and $\sigma^2 = 0.5$	48 pieces

What are degrees of freedom (df)?

Degrees of freedom (df): the remaining allowed ways you could move the data

If we make as many assumptions as there are observations, nothing left to estimate

Significance tests

So far, we calculated t-stats for a mean, to test whether it is different from zero.

We can do this in `R` as follows:

```
> t.test(gdp.dem)

        One Sample t-test

data:  gdp.dem
t = 8.6799, df = 21, p-value = 2.174e-08
alternative hypothesis: true mean is not equal to 0
```

or by "hand":

```
tstat <- mean(gdp.dem)/(sd(gdp.dem)/(length(gdp.dem)^(0.5)))
pval <- 2*pt(abs(tstat),df,lower.tail=FALSE)
```

But we're really more interested in whether the Dem and Rep GDP means are significantly different

So we need a t-test for the *difference* in means.

t-test for comparison of means

As with a single mean, we will calculate a t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{\mathrm{se}(\bar{x} - \bar{y})}$$

then check if the t-statistic exceeds the chosen critical value
or simply calculate the p of seeing so large a t

Because the two samples may have different sizes,
the form of the standard error here is a bit messy:

$$\mathrm{se}(\bar{x} - \bar{y}) = \sqrt{\left(\frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}\right) \times \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

t-test for comparison of means

Unfortunately, the number of degrees of freedom, ν, is now ambiguous, also because the samples could be different sizes

An estimate of the dfs for the comparison of means of different-sized samples is:

$$\hat{\nu} = \frac{\left(\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}\right)^2}{\frac{\hat{\sigma}_x^4}{n_x^2(n_x-1)} + \frac{\hat{\sigma}_y^4}{n_y^2(n_y-1)}}$$

(Don't worry, you'll never need to do this by hand)

t-tests for difference of means in R

It's easy to do this test in R

```
library(stats)
t.test(x = gdp.dem,              #  First group
       y = gdp.rep,              #  Second group
       mu = 0,                   #  Null hypothesis
       conf.level = 0.95         #  Desired confidence level
       )
```

```
t = 2.0366, df = 43.679, p-value = 0.04778
alternative hypothesis: true difference in means is not equal to 0
```

How do we interpret the above, in plain English?

Confidence intervals

Some notation:

$\bar{x}$ The mean of a sample of n values of x

$\bar{x}^{\text{population}}$ The true mean from the population of x; unknown

$\bar{x}^{\text{upper}}$ Upper bound of a confidence interval around $\bar{x}$

$\bar{x}^{\text{lower}}$ Lower bound of a confidence interval around $\bar{x}$

What we'd really like to know:

The (objective) probability that the population mean lies between $\bar{x}^{\text{lower}}$ & $\bar{x}^{\text{upper}}$

We can never know this from an incomplete sample.

No one will ever devise a technique to show us.

Not possible within frequentist inference.

Confidence intervals

This is impossible to find:

The (objective) probability that the population mean lies between $\bar{x}^{\text{lower}}$ & $\bar{x}^{\text{upper}}$

Now what?

Confidence intervals: Option 1

1. Drop the word "objective", and calculate a subjective probability.
   This is the method used by Bayesian inference:

   Based on an initial, subjective assessment (e.g., personal reading of past research), define a prior distribution of $\bar{x}^{\text{population}}$, $P(\bar{x}^{\text{population}})$.

   One way to read this prior distribution:
   A set of intervals within which you believe $\bar{x}^{\text{population}}$ lies with $1 - \alpha$ probability.

   Thus we already have (by assumption alone) an answer to our question. But we want to update that subjective probability interval to account for inference from our data.

   We update our distibution of $\bar{x}^{\text{population}}$ by Bayes rule:

   $$P(\bar{x}^{\text{population}}|\bar{x}) = P(\bar{x}^{\text{population}})\frac{P(\bar{x}|\bar{x}^{\text{population}})}{P(\bar{x})}$$

   Now we have a "posterior" set of $1 - \alpha$ Bayesian credible intervals implied by $P(\bar{x}^{\text{population}}|\bar{x})$.

   These intervals give the probability $\bar{x}^{\text{population}}$ is between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$, based on the data and our prior beliefs.

Confidence intervals: Option 2

2. If we want to remain "objective" – that is, ignore our prior beliefs – we cannot calculate a probability that $\bar{x}^{\text{population}}$ lies between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$.

The frequentist solution:

A new concept, called **confidence**, used in place of probability.

We are 95% **confident** that $\bar{x}^{\text{population}}$ lies between $\bar{x}^{\text{lower}}$ and $\bar{x}^{\text{upper}}$ when

in repeated samples from the population, 95% of intervals constructed in such a fashion contain the truth $\bar{x}^{\text{population}}$

Why isn't this a probability?

It's a statement about the *asymptotic* properties of the t distribution,
not the data we observed.

We just don't know, based on the draw we made, whether the truth lies inside the frequentist confidence interval. And we don't know the probability this is true

Confidence intervals for means

To get the $100(1-\alpha)$% confidence interval for the mean of x,

$$\begin{aligned}\bar{x}^{\text{lower}} &= \bar{x} - t_{\alpha/2,n-2} \times \hat{\sigma}_x \\ \bar{x}^{\text{upper}} &= \bar{x} + t_{\alpha/2,n-2} \times \hat{\sigma}_x\end{aligned}$$

where $t_{\alpha/2,n-2}$ is the critical value of the t distribution with $n-2$ degrees of freedom and a probability of $a/2$ to the right

"We're 95% confident $\bar{x}^{\text{true}}$ is equal to $\bar{x}$, $\pm$ a few standard deviations of x"

Plus or minus how many std devs? $t_{\alpha/2,n-2}$ many.

For a large dataset, the 95% CI will tend to be $\pm 1.96\hat{\sigma}$

For a large dataset, the 99% CI will tend to be $\pm 2.576\hat{\sigma}$

But you need to calculate the exact value using `qt(alpha/2,n-2)`

Confidence intervals for difference of means

To get the $100(1-\alpha)$% confidence interval for a difference of means,

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\hat{\nu}} \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

where $t_{\alpha/2,\hat{\nu}}$ critical value of the t distribution with $\hat{\nu}$ degrees of freedom and a probability of $a/2$ to the right

$\hat{\nu}$ is estimated as before

Just leave this one up to `R`. . .

Confidence intervals for means in `R`

The `t-test()` command provides CIs as well:

```
library(stats)
t.test(x = gdp.dem,                #  First group
       y = gdp.rep,                #  Second group
       mu = 0,                     #  Null hypothesis
       conf.level = 0.95           #  Desired confidence level
       )

t = 2.0366, df = 43.679, p-value = 0.04778
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 0.01400377    2.72306582

sample estimates:
mean of x    mean of y
 3.094356     1.725821
```

Confidence intervals

How do we report a confidence interval?

Democratic presidents enjoyed growth rates 1.37 points higher [95% CI: 0.01 to 2.72] than their Republican counterparts.

or

Democrats enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

We could calculate any CI we wish: 90 percent, 80 percent, 50 percent, etc.

The most commonly used are: 90, 95, and 99.

Can we get 100 percent CIs?

Not unless we can *logically* reject values outside that interval

Significance tests vs confidence intervals

Problems with significance tests & confidence intervals:

- Simply a commitment to a certain error rate, given all assumptions met
- Suppose null is really true. But eventually, someone will find a "significant" difference from any null by chance. If you publish all (and only) significant results, journals will be full of falsehoods, and file drawers of rejected papers full of truth.

This is known as the *file drawer problem*

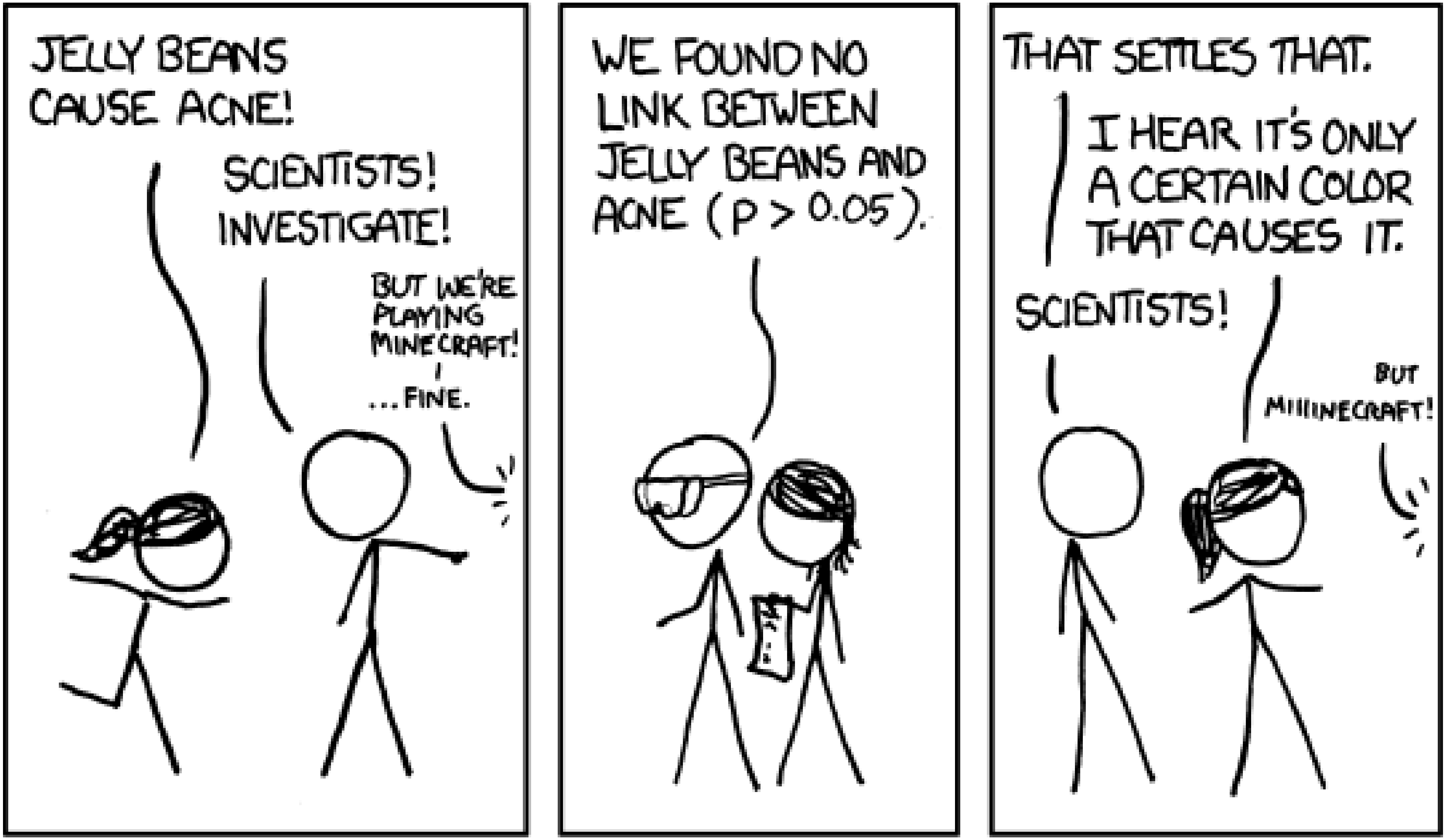
"Fishing." A research "technique" in which the researcher

1. modifies an insignificant model until a significant result occurs, (regardless of one's prior beliefs of what models are likely correct)

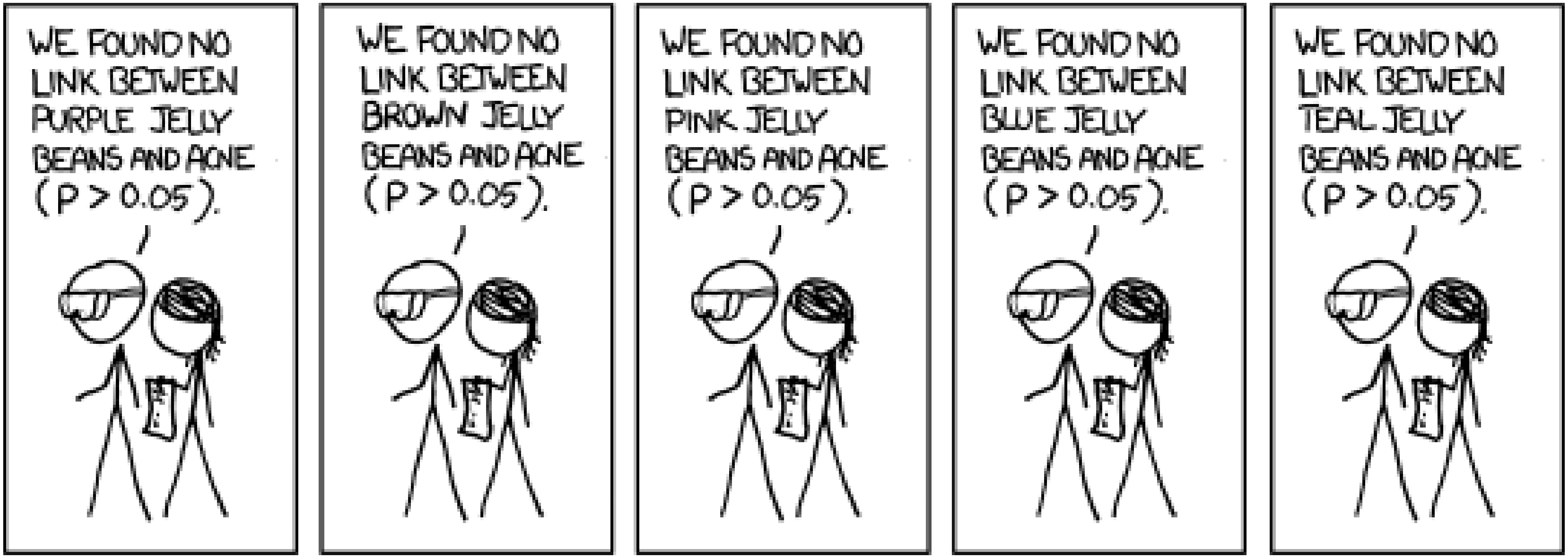
2. Publishes only the final ("significant") result

What t-statistic do you expect a "fisher" to end up with / not end up with?

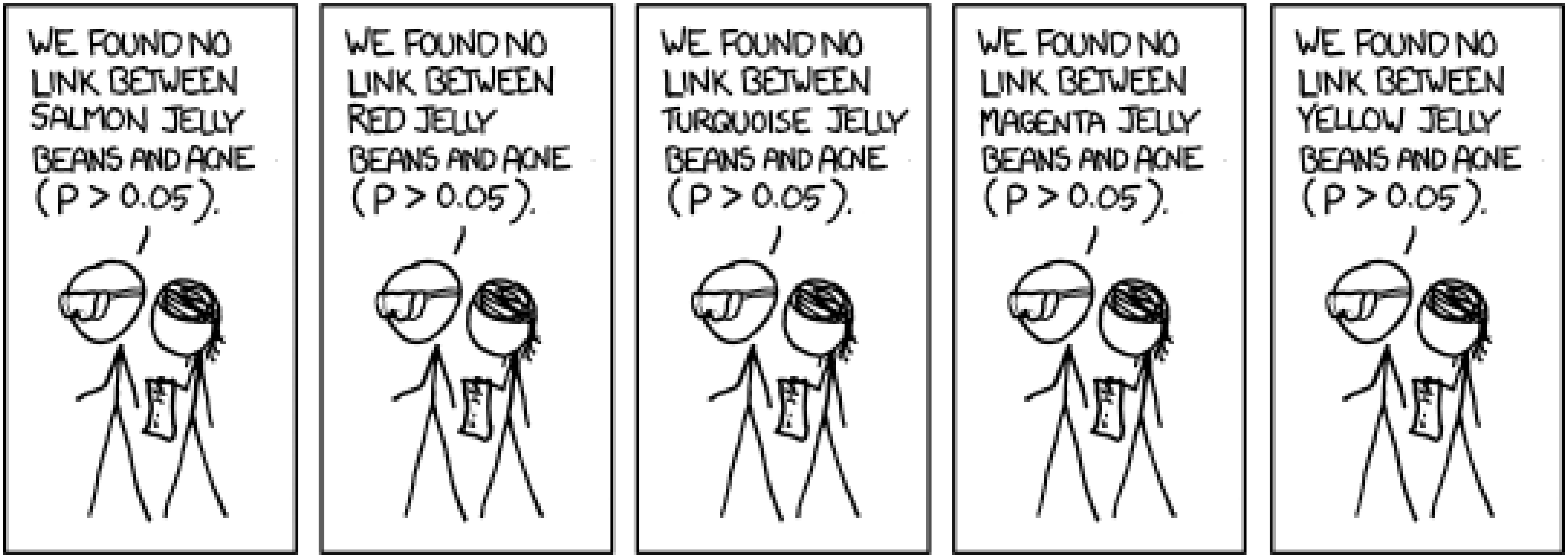
What happens when the next researcher tries to replicate this result in a new sample?


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

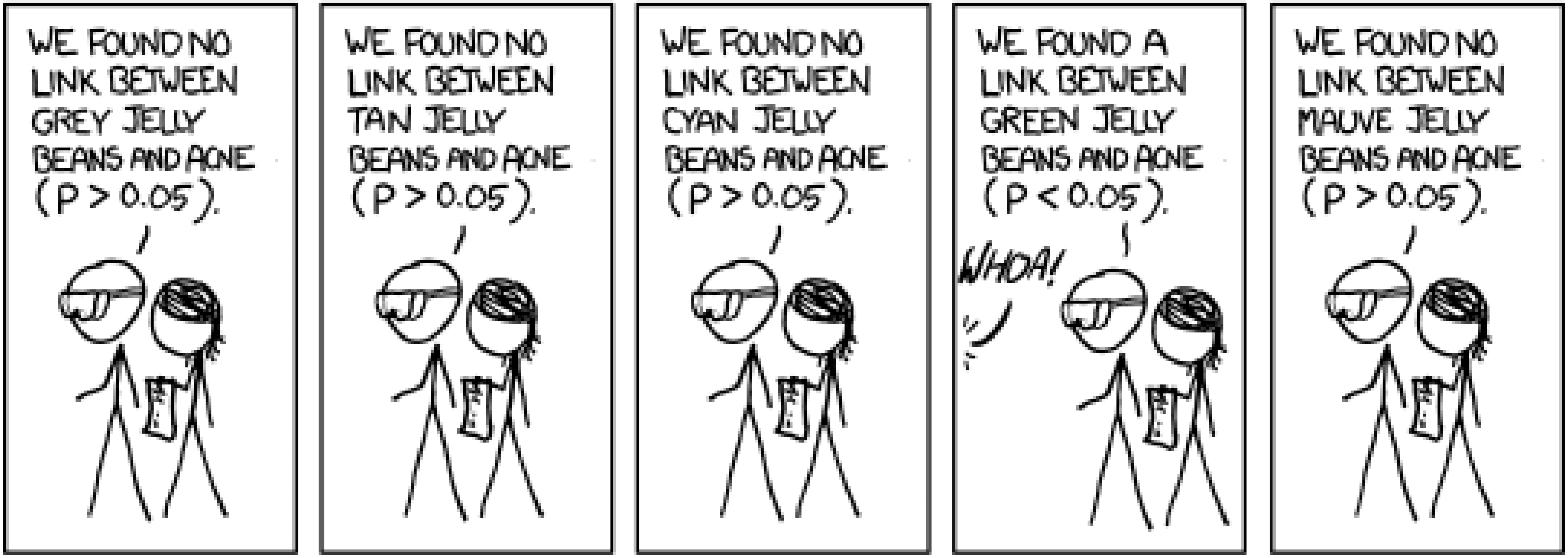
xkcd.com/882


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

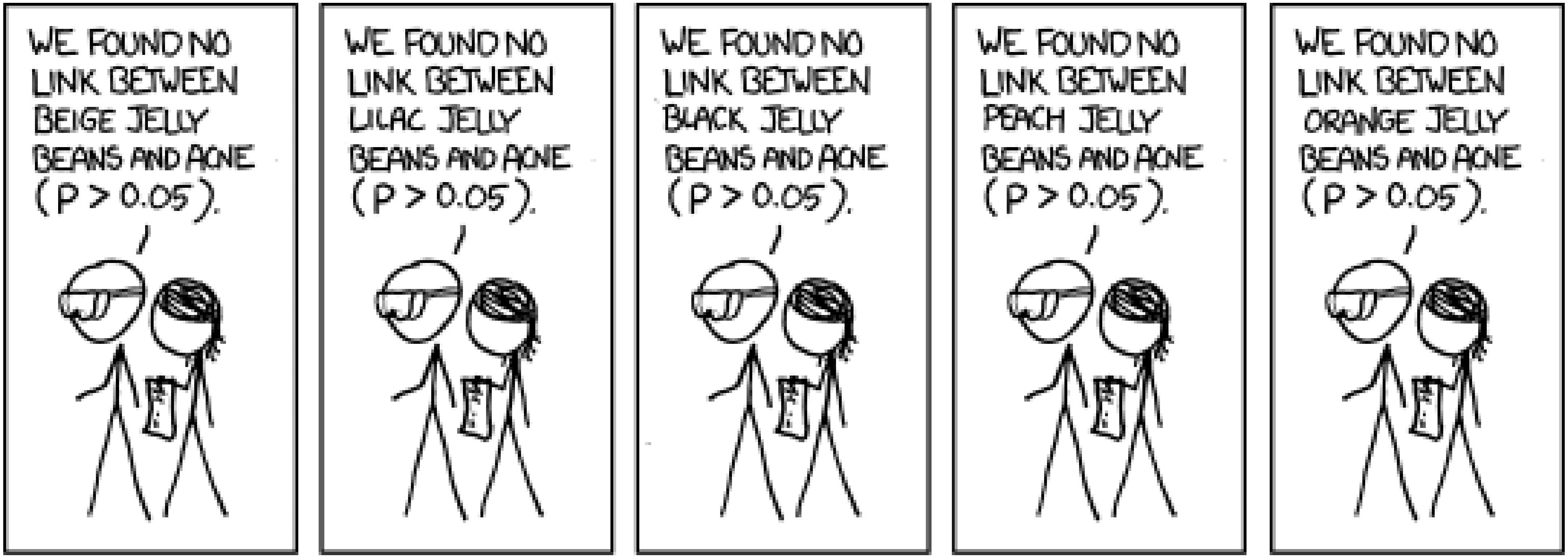
xkcd.com/882


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

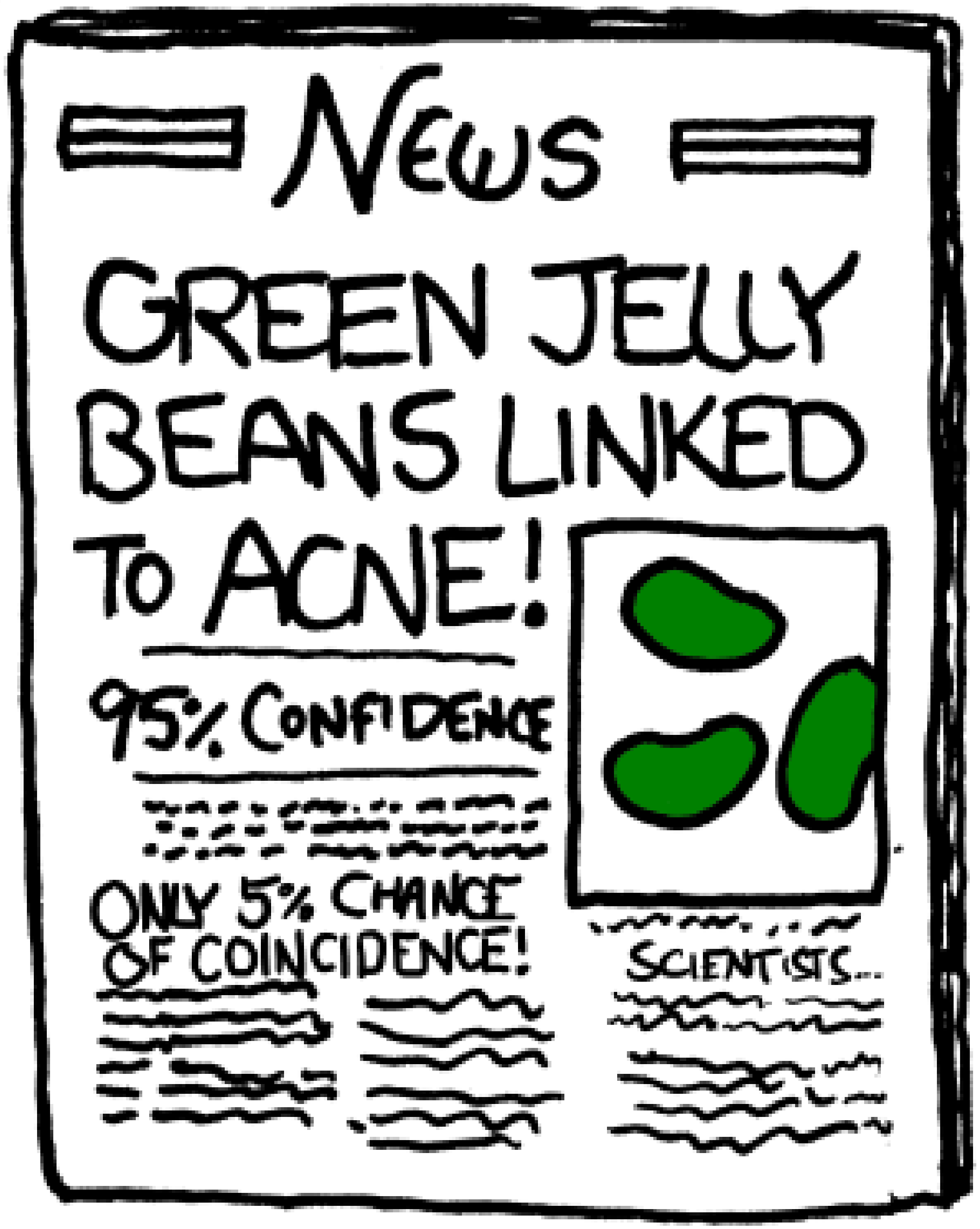
xkcd.com/882


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

xkcd.com/882


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

xkcd.com/882



NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com/882

Significance tests vs confidence intervals

Problems with **both** significance tests & confidence intervals:

- Simply a commitment to a certain error rate, given all assumptions met
- File drawer problem and fishing as a research agenda

Problems with significance tests that CIs overcome:

- Hypothesis tests are "weak"
  in the sense that they don't tell you if the alternative hypothesis (your finding)
  was different from the new null hypothesis $X^{\text{null}} + 0.000001$
- p-values encourage a focus on statistical significance
  rather than substantive significance (the original point of the model!)
  Avoid "star-gazing," or "measuring" effects by p-values

Significance tests vs confidence intervals

Confidence intervals are not perfect

Share many of the same limits & awkwardness of significance tests

And people tend to mistake them for probability intervals

But which would you rather say or read:

- Compared to Republicans, the effect of Democratic presidents on the economy is significantly positive at the 0.05 level.
- Democratic presidents enjoyed 1.37 points higher growth than Republicans, with a 95 percent confidence interval of 0.01 to 2.72.

Even with the ambiguity of the words "significance" and "confidence,"
the latter says more and says it more clearly,

and avoids conflation of statistical and substantive significance

Frequentism versus Bayes

Frequentist confidence intervals harder to interpret than Bayesian confidence intervals

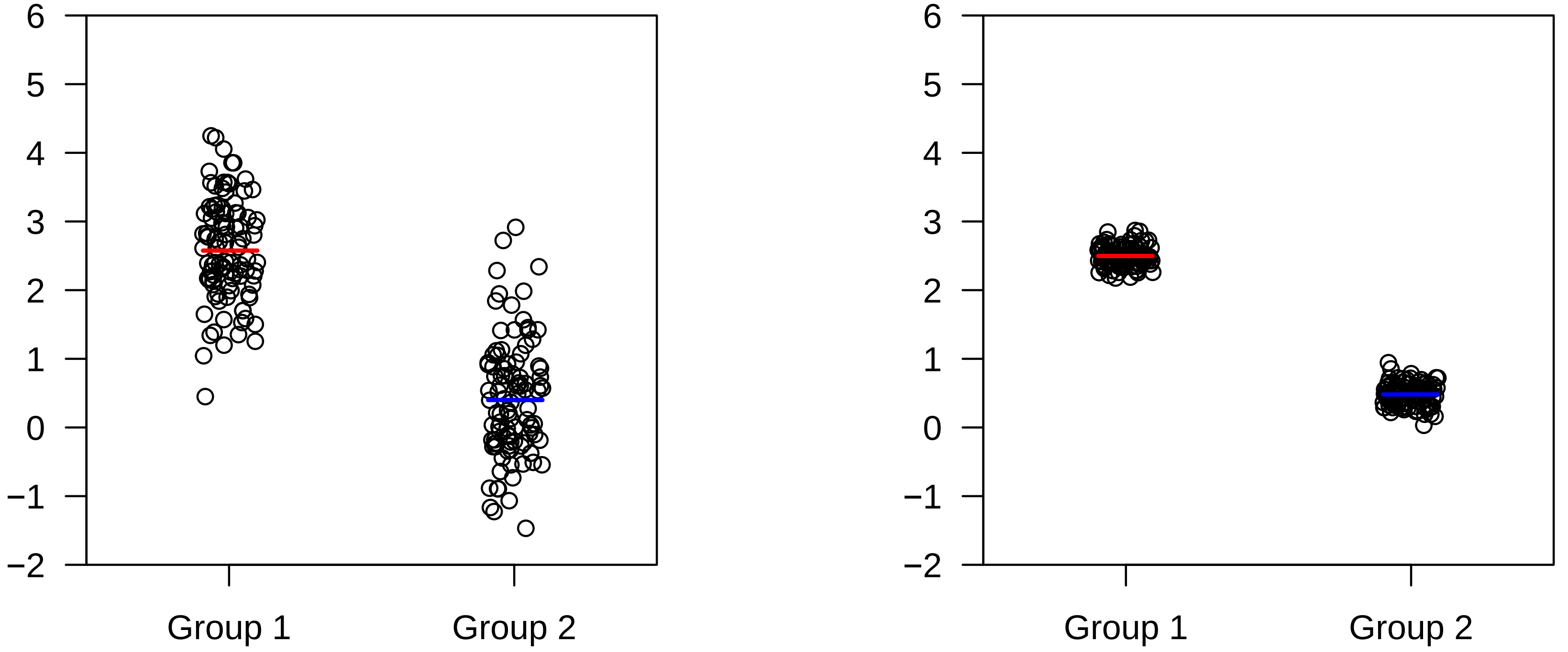
Bayesian confidence intervals' clear meaning comes at a price:

- Subjective priors (but some would argue these are a benefit)
- Mathematical & computational complexity

Bayesian inference is a topic for more advanced classes

UW one of best places to learn Bayesian methods (CSSS)

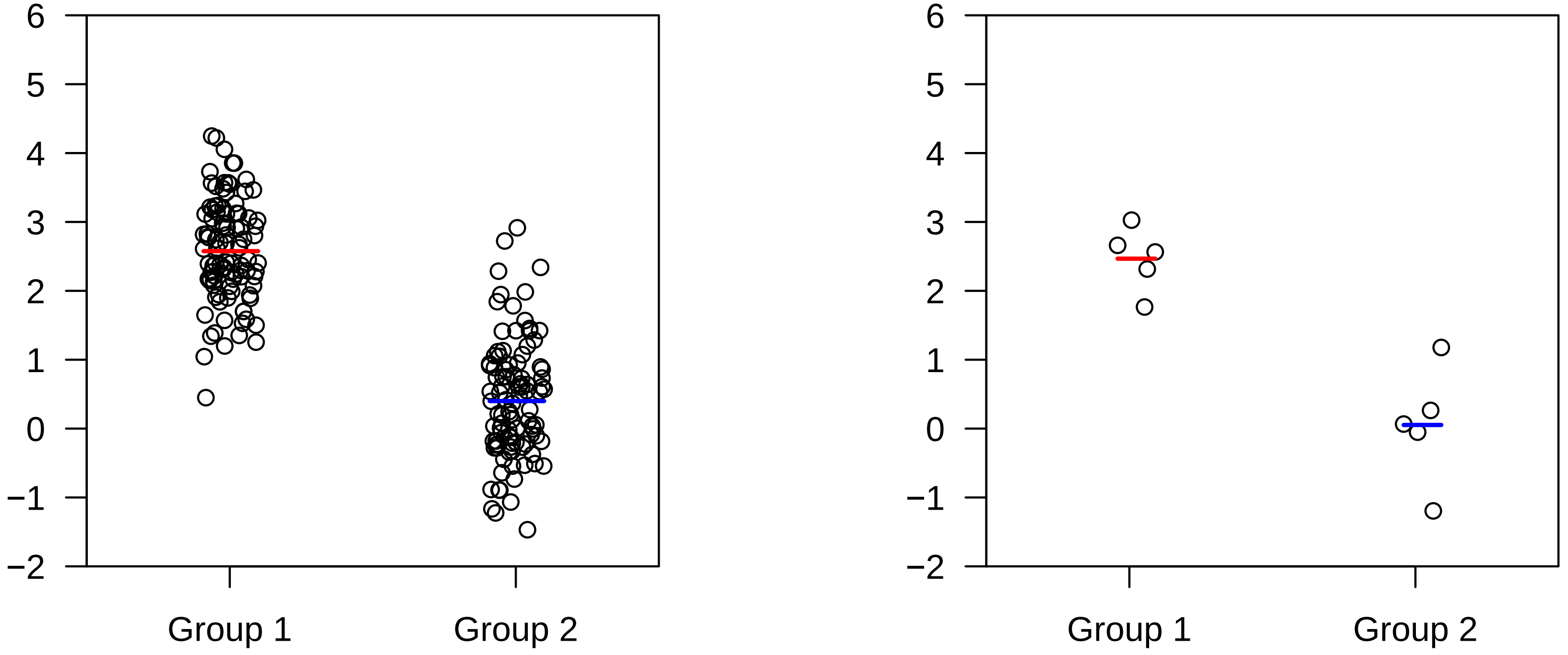
“These means are different.” Equally confident?



Same means in each pair; different variances:
In order, $\mathcal{N}(2.5, 0.72)$, $\mathcal{N}(0.5, 0.72)$, $\mathcal{N}(2.5, 0.02)$, $\mathcal{N}(0.5, 0.02)$

Lower variance in this case gives more info per observation (↑ signal to noise)

“These means are different.” Equally confident?



Truth for both plots the same: G1 from $\mathcal{N}(2.5, 0.72)$, G2 from $\mathcal{N}(0.5, 0.72)$

Small samples lower significance. But may *look* like very strong relationships. Empirically, science journals seem overly optimistic of statistical significance of small N results, and are more likely to publish them (file-drawer problem again)

Inference for regression

We can calculate confidence intervals and significance tests for any estimate, including regression coefficients

Regression coefficients are random variables

What is their sampling distribution?

$\rightarrow$ Distribution of $\hat{\beta}_1$ given repeated sampling of x, y from the population

In practice, we sample x, y once and then run the regression.

But imagine we did in thousands of times.

We'd pile up the estimated $\hat{\beta}_1$, and get a histogram.

What is the mean and variance of this distribution?

Inference for regression

What is the mean and variance of the distribution of $\hat{\beta}_1$?

The mean is the expected value of $\hat{\beta}_1$.

If:

- the relationship between y and x really is linear,
- and x and ε really are independent,

then we expect $\hat{\beta}_1$ to match the truth, on average:

$$\mathrm{E}(\hat{\beta}_1) = \beta_1$$

Inference for regression

If:

- the relationship between y and x really is linear,
- and x and ε really are independent,
- *and* y really is iid Normal conditional on x,

the variance of $\hat{\beta}_1$ will be

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Upshot: Can estimate the variance of $\hat{\beta}_1$ with a single draw from its distribution

This is nothing short of magical

Inference for regression

We've already seen the variance-covariance matrix of the disturbances $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$

To estimate the variance-covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta}$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{\Sigma}}_{\boldsymbol{\beta}} = \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right]$$

Note that:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}
\end{aligned}$$

Inference for regression

Substituting, we find

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} &= \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] \\
&= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2_{\varepsilon}\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2_{\varepsilon}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This form is analogous to the bivariate version above, with two differences:

It calculates every standard error at once

It calculates *covariances*, which allow the estimates $\hat{\beta}_i$ and $\hat{\beta}_j$ to be correlated

Inference for regression

Putting these together, and assuming $\mathrm{E}(y|x)$ really is Normal, we have the distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \quad \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

or, in matrix form, the joint distribution of all the $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\boldsymbol{\beta}, \quad \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

Even if $\mathrm{E}(y|x)$ isn't Normal, this holds approximately as $n \to \infty$

Recall:
$\mathcal{MVN}(\cdot,\cdot)$ represents the Multivariate Normal distribution, which allows several variables to be both Normal and potentially correlated with each other

Inference for regression

We now have an estimate of the standard deviation of $\hat{\beta}_1$.

This is called the standard error of β_1:

$$\mathrm{se}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

This is the second column of regression output in `R`.

The se's can also be found by taking the square roots of the diagonal elements of the variance-covariance matrix:

$$\mathrm{se}(\hat{\boldsymbol{\beta}}) = \sqrt{\mathrm{diag}\left(\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right)}$$

Notice that standard errors get bigger (less precise) when the data get noiser,
or when the x's have little variation

Inference for regression

Assuming the β's are normally distributed, and their standard errors are χ^2 distributed, we can construct the following t-distributed test statistic

$$t = \frac{\hat{\beta}_1 - \beta_1^{\text{null}}}{\text{se}(\beta_1)}$$

commonly known as the t-statistic.

It is distributed t with $n - k - 1$ degrees of freedom

To conduct significance tests, just calculate the p-value

That is, the area under the tails beyond $(\hat{\beta}_1 - \beta_1^{\text{null}})/\text{se}(\beta_1)$ of a t-distribution with $n - k - 1$ dfs

In R: `2*(1-pt(tstat,df))`

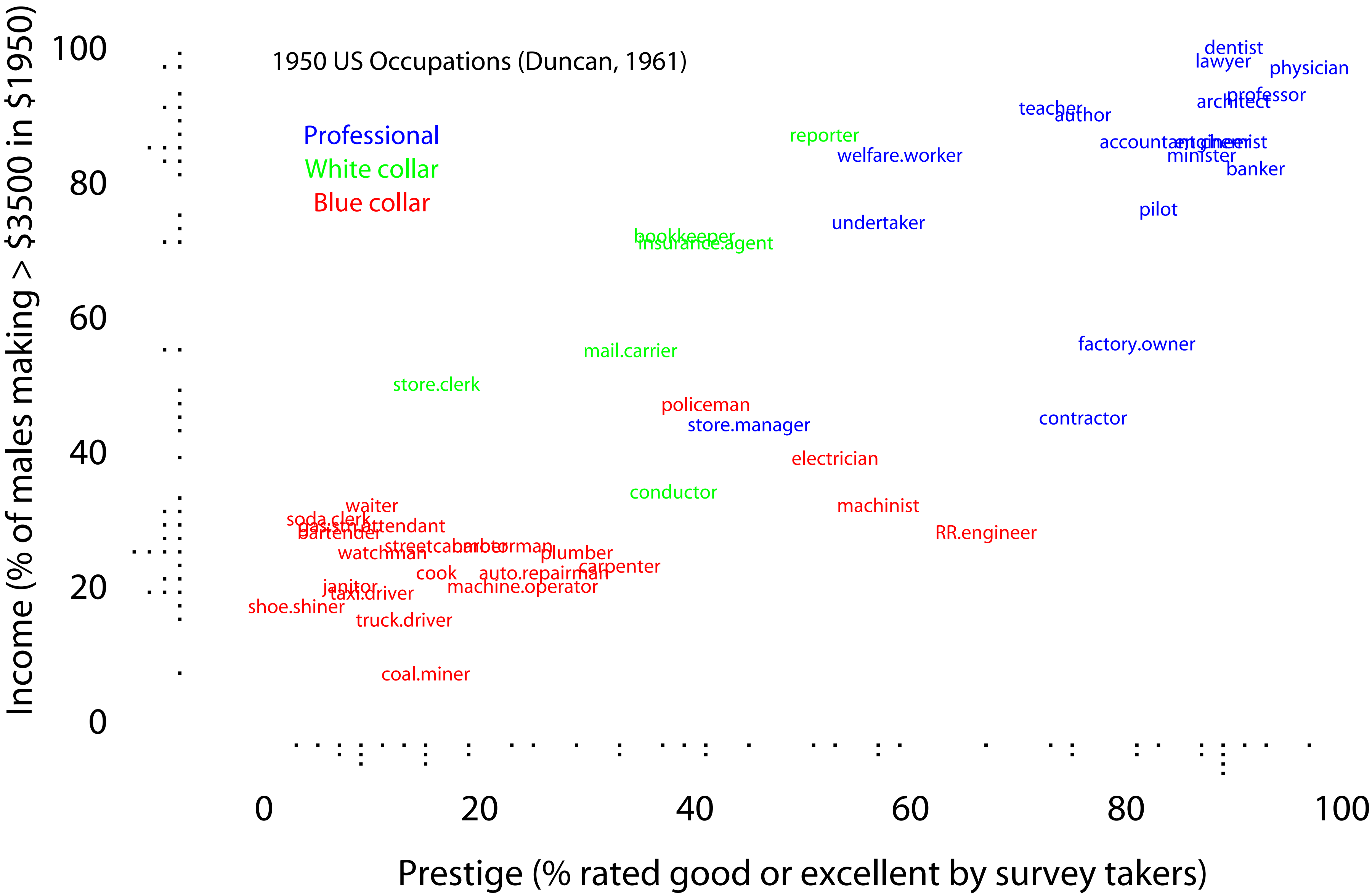
Example: Occupational Prestige & Income

Classic data from sociology. Three variables

- Prestige of occupations, as rated by surveys
- Income of occupations (averaged across males)
- Type of occupation (blue collar, white collar, professional)

Data is in `R`.

Load the `car` library and run `data(Duncan)` and `help(Duncan)`


1950 US Occupations (Duncan, 1961)
Professional
White collar
Blue collar
Income (% of males making > $3500 in $1950)
100
80
60
40
20
0
Prestige (% rated good or excellent by survey takers)
0
20
40
60
80
100
dentist
lawyer
physician
architect
professor
teacher
author
reporter
welfare.worker
accountant
engineer
chemist
minister
banker
pilot
undertaker
bookkeeper
insurance.agent
mail.carrier
factory.owner
store.clerk
policeman
store.manager
contractor
electrician
conductor
waiter
machinist
soda.clerk
gas.stn.attendant
bartender
RR.engineer
streetcar.motorman
barber
watchman
plumber
carpenter
cook
auto.repairman
janitor
machine.operator
taxi.driver
shoe.shiner
truck.driver
coal.miner

```
> model1 <- prestige~income+education
> lm.res1 <- lm(model1, data=Duncan)
> summary(lm.res1)

Call:
lm(formula = prestige ~ income + education)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5380  -6.4174   0.6546   6.6051  34.6412

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05 ***
education    0.54583    0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-Squared: 0.8282,     Adjusted R-squared:  0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

To find the t-statistics & p-values, use the `summary()` command.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05 ***
education    0.54583    0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note `1.05e-05` $= 0.0000105$

Or, you could calculate yourself:

```
model1 <- prestige~income+education     # model specification
lm.res1 <- lm(model1, data=Duncan)      # run the regression
beta <- lm.res1$coefficients            # retrive the of betas
vc <- vcov(lm.res1)                     # retrieve the var-cov matrix
se <- sqrt(diag(vc))                    # calc a vector of ses
tstat <- beta/se                        # calc vector of tstats
pval <- 2*(1-pt(tstat,42))              # calc p-values
```

Confidence intervals for regression coefficients

Standard errors, t-tests, and p-values take expertise to read

They are also subject to misinterpretation

(E.g., smaller p-values do *not* imply a bigger substantive effect)

CIs turn the standard errors into something more people can understand

To get the $100(1-\alpha)$% confidence interval for $\hat{\beta}_1$,

$$\begin{aligned}
\hat{\beta}_1^{\text{lower}} &= \hat{\beta}_1 - t_{\alpha/2,n-k-1} \times \hat{\sigma}_{\hat{\beta}_1} \\
\hat{\beta}_1^{\text{upper}} &= \hat{\beta}_1 + t_{\alpha/2,n-k-1} \times \hat{\sigma}_{\hat{\beta}_1}
\end{aligned}$$

Confidence intervals for regression coefficients

How to calculate CIs for coefficients in `R`

By hand:

```
lower.95 <- beta - qt(0.025,42)*se
upper.95 <- beta + qt(0.025,42)*se
```

Why are we using `qt`? Why `0.025`?

The easy way:

```
library(stats)
confint(lm.res1, level=0.95)

                  2.5 %    97.5 %
(Intercept) -14.6857892 2.5564634
education     0.3475521 0.7441158
income        0.3572343 0.8402313
```

Confidence intervals for regression coefficients

Using confidence intervals, we can improve the initial summary table:

		95% Conf Interval	
Variable	Estimate	Lower	Upper
Income	0.60	[0.36,	0.84]
Education	0.55	[0.38,	0.74]
Intercept	-6.06	[-14.69,	2.46]
N	45		
s.e.r.	13.4	(this is $\hat{\sigma}_{\varepsilon}$)	
R^2	0.83	(this line optional)	

Table 1: Determinants of occupational prestige. Entries are linear regression parameters and their 95 percent confidence intervals.

Think about everything you put in these tables:

- what readers need to see to fully understand your results
- what superfluous R output you can delete
- how to make the results clear for as large an audience as possible

Substantive & statistical significance

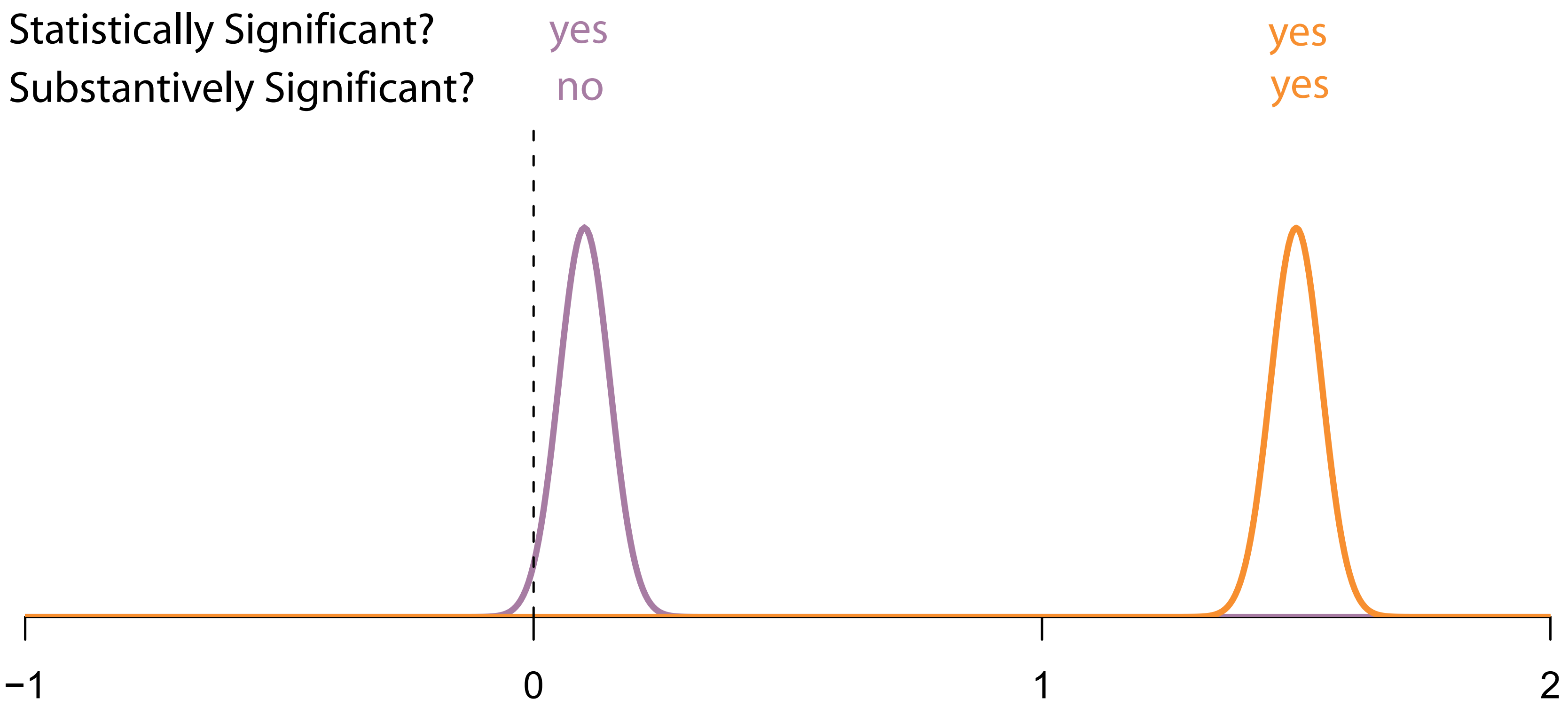
Don't over interpret p-values

They only show statistical significance

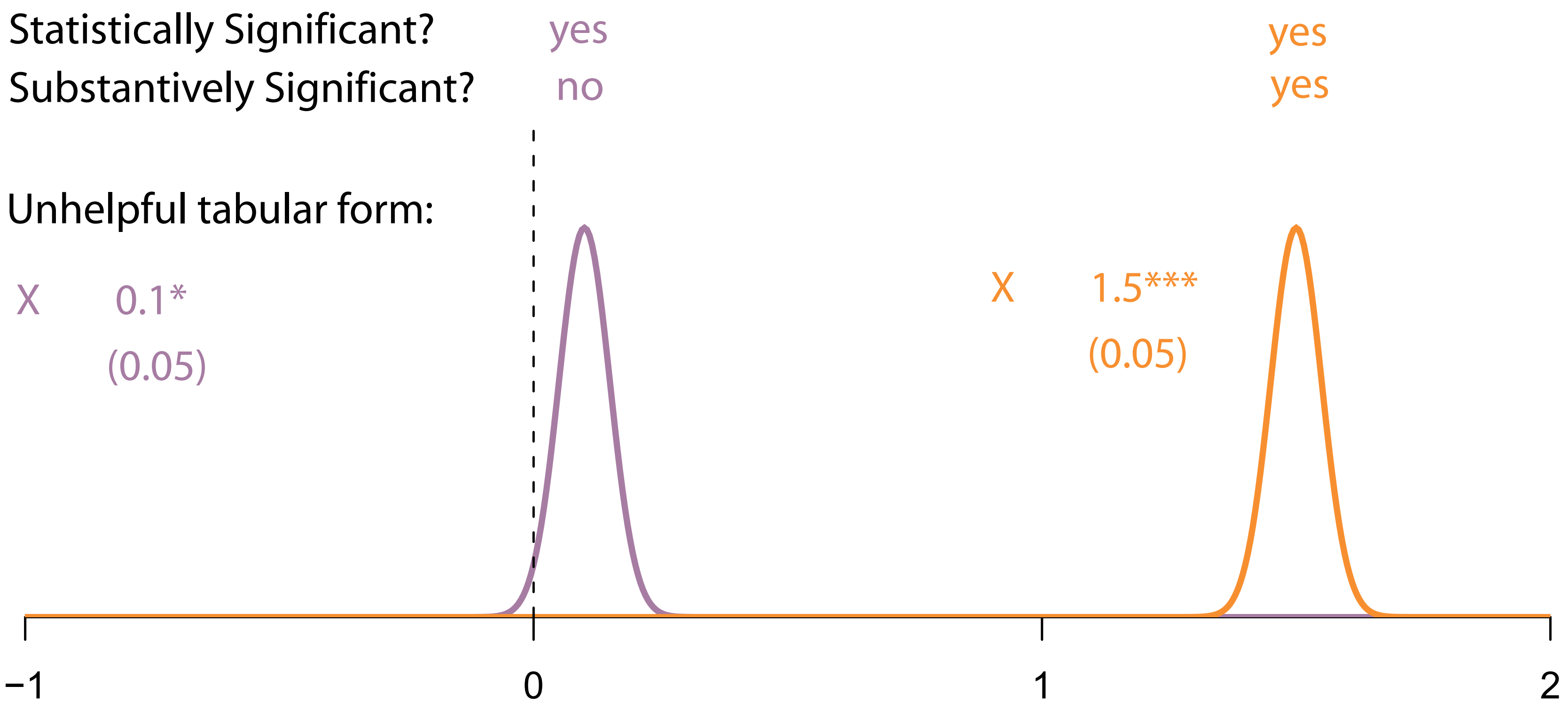
Statistical and substantive significance can interact

A look at some hypothetical distributions of $\hat{\beta}_1$ helps frame the possibilities

Perils of stargazing



Perils of stargazing



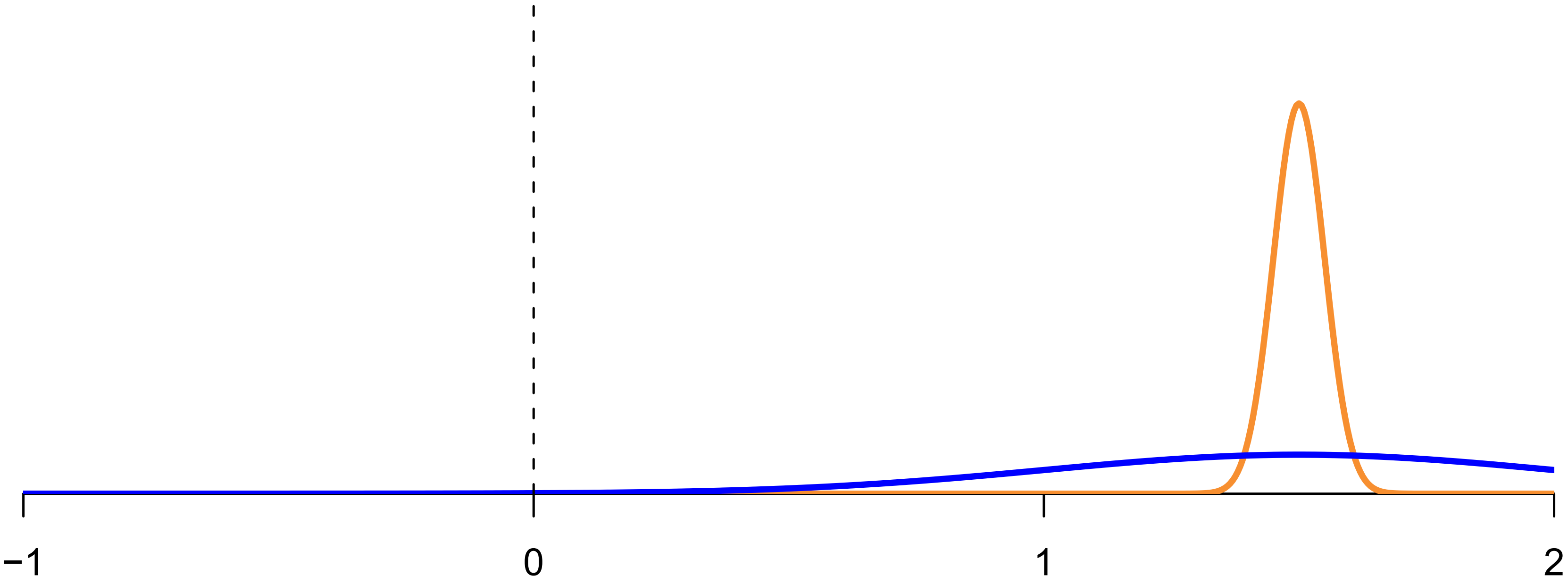
These estimated β's will both be starred in regression output.

Often, only the estimate to the right will be significant in a substantive sense

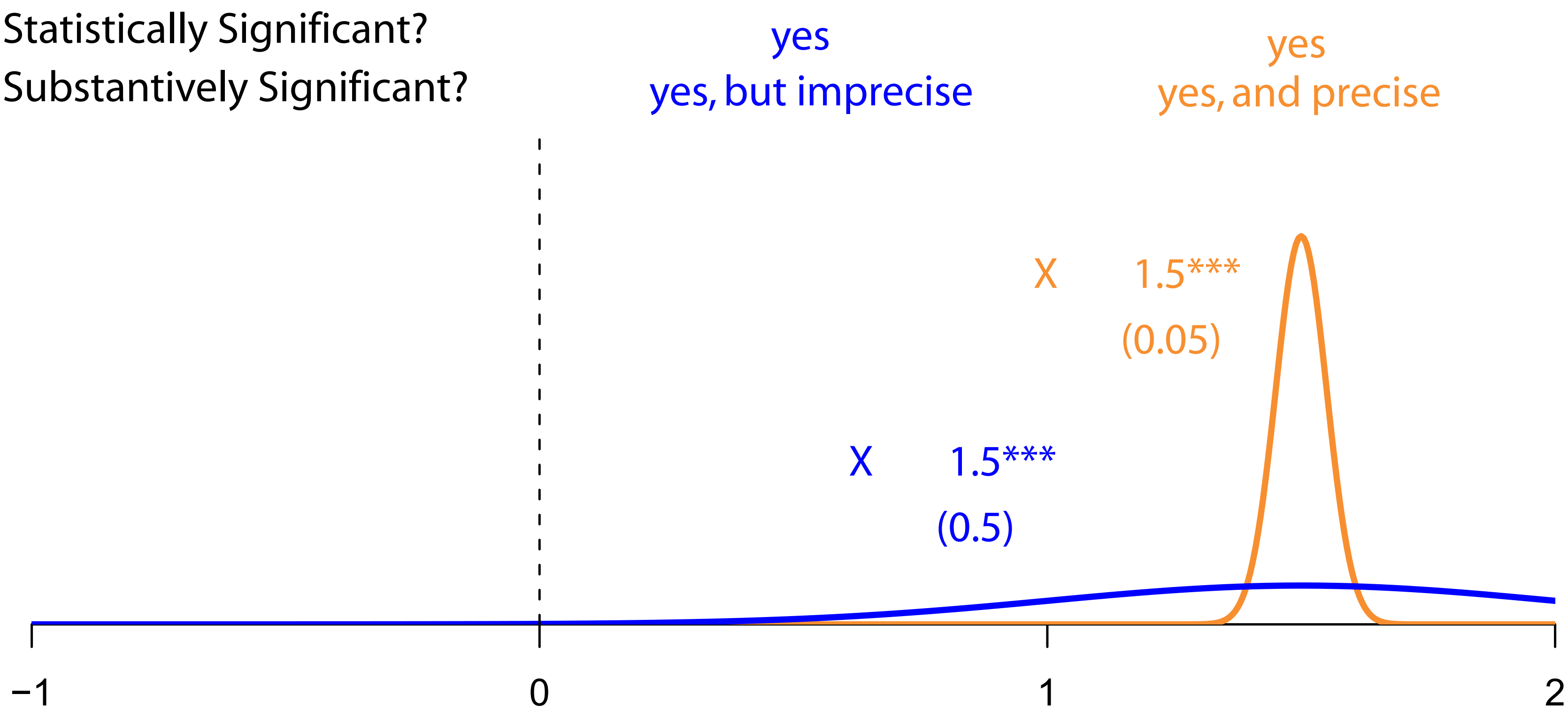
The estimate on the left is a precise zero

Perils of stargazing

Statistically Significant?	yes	yes
Substantively Significant?	yes, but imprecise	yes, and precise



Perils of stargazing



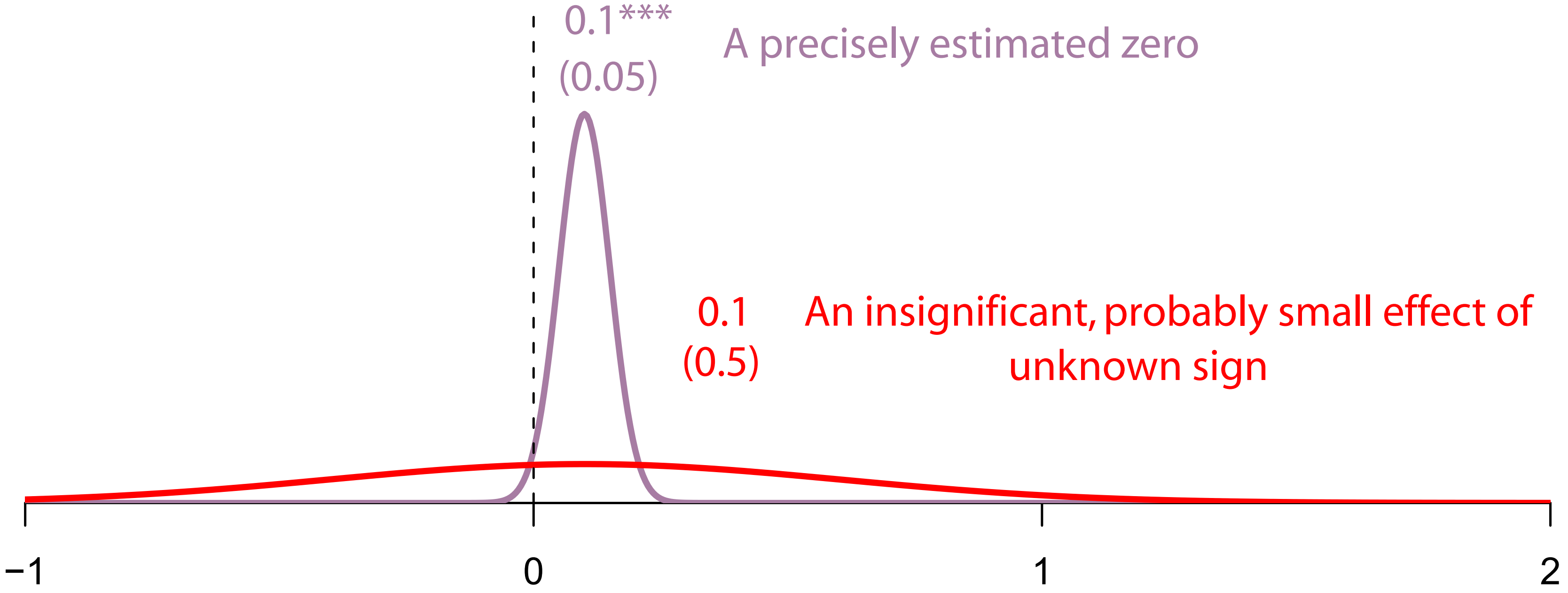
These estimated β's will both be heavily starred in regression output.

They are both substantively significant as well, with identical point estimates

But the orange curve is much more precisely estimated

The blue estimate may be much smaller or larger. Best shown with a CI

Perils of stargazing



How do you verify a null effect? Precise zeros

Sometimes, researchers mistake the precise zero for a positive effect

Confidence interval for expected values

We can calculate the CIs around $\hat{Y}$ as well.

For example, what is the 95% CI around $\widehat{\text{Prestige}}_c$ in:

$$\widehat{\text{Prestige}}_c = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}_c + \hat{\beta}_2 \text{Education}_c$$

The uncertainty in each estimate will “combine” to form the uncertainty in $\widehat{\text{Prestige}}_c$.

In this example,

$$\begin{array}{cccccc}
\widehat{\text{Prestige}}_c & = & -6.1 & 0.60 \times \text{Income}_c & + & 0.55 \times \text{Education}_c \\
 & & [-14.7, 2.6] & [0.36, 0.84] & & [0.35, 0.74] \\
 & & & & & \\
47.7 & = & -6.1 & 0.60 \times 41.9 & + & 0.55 \times 52.6 \\
[43.7, 51.7] & & [-14.7, 2.6] & [0.36, 0.84] & & [0.35, 0.74]
\end{array}$$

In words, when income and education are held at their means,
we expect that prestige will equal 47.7 with a 95 % CI of 43.7 to 51.7.

Confidence interval for expected values

How do we calculate confidence intervals around $\hat{y}$ in `R`?

1. Estimate the model

2. Choose hypothetical values of the covariate at which you want to calculate $\hat{y}$ and its CI.

3. Use the `predict()` function to obtain the expected y and its CI

Some examples:

```
# To get CIs around all the fitted values
res <- lm(y~x+z)
pred <- predict(res, interval="confidence", level=0.95)
yhat <- pred[,1]
yhat.lower <- pred[,2]
yhat.upper <- pred[,3]
```

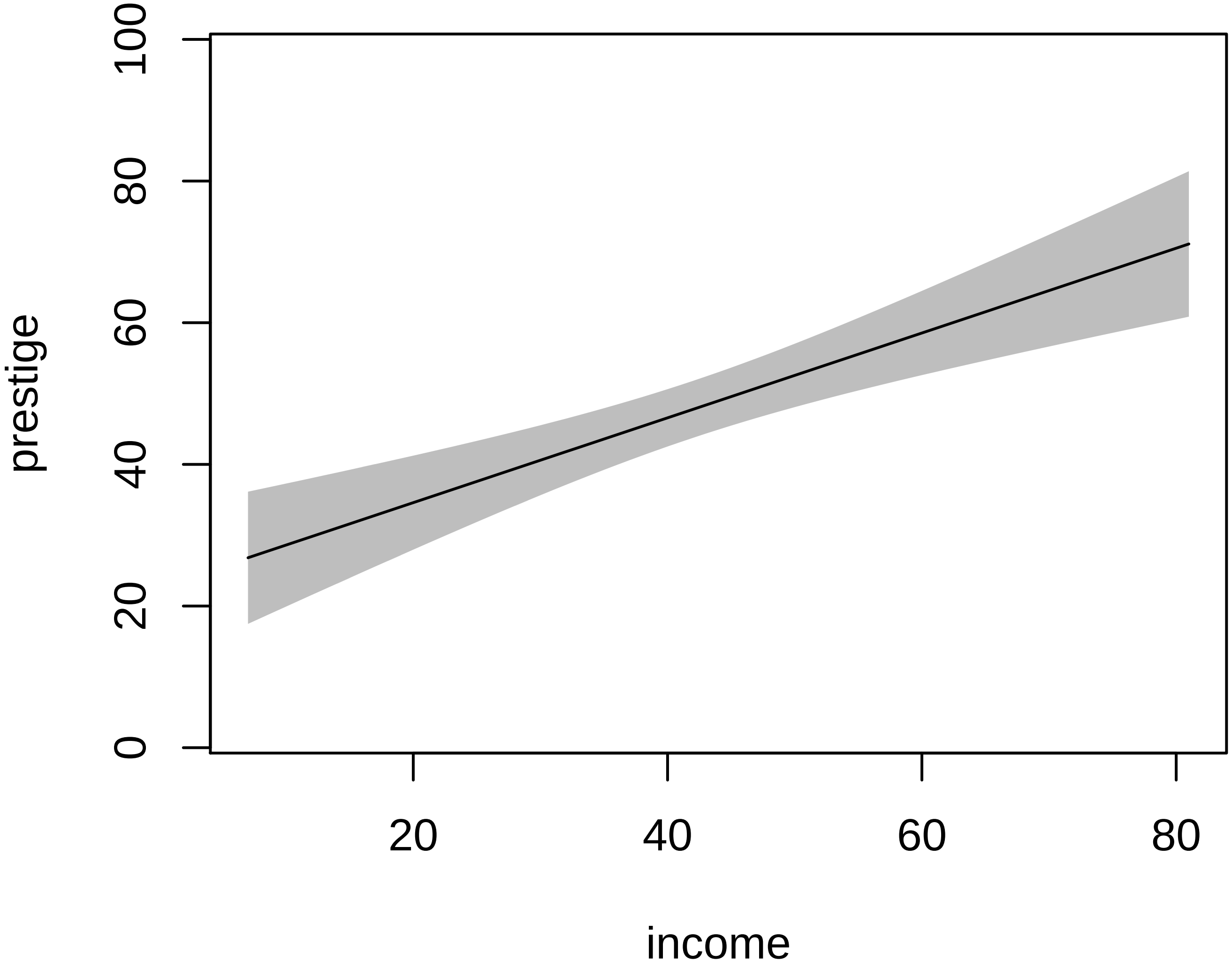
Confidence interval for expected values

```
# To get CIs for yhat given a set of hypothetical x & z values
res <- lm(y˜x+z)
xhyp <- seq(min(x), max(x), 0.01)
zhyp <- rep(mean(z), length(xhyp))
hypo <- data.frame(x=xhyp, z=zhyp)
pred <- predict(res,
                newdata=hypo,
                interval="confidence",
                level=0.95)
yhat <- pred[,1]
yhat.lower <- pred[,2]
yhat.upper <- pred[,3]
```

The code above is very useful for adding confidence intervals to a plot.

We can run through a sequence of possible x values, holding z constant,
and predict y and its confidence interval,
then plot the confidence interval as an envelope around y

Confidence interval for expected values



Confidence interval for expected values

Interpretation:
All we can say with 95 percent confidence is that the line
– the relation b/w prestige and income –
lies in this envelope

Very useful to show, especially if the relationship is curved in some way

It's easy to plot CIs with dashed lines. . .

```
lines(x=xhyp, y=yhat.lower, lty="dashed")
lines(x=xhyp, y=yhat.upper, lty="dashed")
```

but I prefer shaded regions to dotted lines. (lots of lines gets confusing)

You can make shaded regions using the `polygon()` command

Just be sure to plot the polygon before you add any points or lines,
so it shows up behind the points and lines, instead of covering them up

Complete code for above figure

```
# Load the occupation data
library(car)
data(Duncan)
attach(Duncan)

# A regression analysis, with inference
model1 <- prestige~income+education     # model specification
lm.res1 <- lm(model1, data=Duncan)      # run the regression
beta <- lm.res1$coefficients             # retrive the of betas
vc <- vcov(lm.res1)                      # retrieve the var-cov matrix
se <- sqrt(diag(vc))                    # calc a vector of ses
tstat <- beta/se                        # calc vector of tstats
pval <- 2*(1-pt(tstat,42))              # calc p-values

# Confidence intervals of betas by hand
lower.95 <- beta - qt(0.025,42)*se
upper.95 <- beta + qt(0.025,42)*se

# Confidence intervals of betas the easy way
library(stats)
confint(lm.res1, level=0.95)
```

```
## Make a plot...
## CIs for yhat given a set of hypothetical income & education
xhyp <- seq(min(income), max(income),1)
zhyp <- rep(mean(education), length(xhyp))
hypo <- data.frame(income=xhyp, education=zhyp)
pred <- predict(lm.res1,
                newdata=hypo,
                interval="confidence",
                level=0.95)
yhat <- pred[,1]
yhat.lower <- pred[,2]
yhat.upper <- pred[,3]

# Start the plot

# Uncomment the below for pdf output (step 1)
#pdf("yhatexample.pdf", width=5, height=4.5)

plot(y=prestige, x=income, type="n")

# Make the x-coord of a confidence envelope polygon
```

```
xpoly <- c(xhyp,
           rev(xhyp),
           xhyp[1])

# Make the y-coord of a confidence envelope polygon
ypoly <- c(yhat.lower,
           rev(yhat.upper),
           yhat.lower[1])

# Choose the color of the polygon
col <- "gray"

# Plot the polygon first, before the points & lines
polygon(x=xpoly,
        y=ypoly,
        col=col,
        border=FALSE
        )

# Plot the fitted line
lines(x=xhyp, y=yhat)
```

```
# Uncomment the below for pdf output (step 2)
#dev.off()
```